Cognitive, Social and Environmental Sources of Bias in Clinical Performance Ratings

Reed G. Williams
Department of Surgery
Southern Illinois University School of Medicine
Springfield, Illinois USA

Debra A. Klamen
Department of Psychiatry
University of Illinois at Chicago
Chicago, Illinois USA

William C. McGaghie
Office of Medical Education and Faculty Development
Northwestern University
Feinberg School of Medicine
Chicago, Illinois USA

Background: Global ratings based on observing convenience samples of clinical performance in vivo form the primary basis for appraising the clinical competence of medical students, residents, and practicing physicians. This review explores cognitive, social, and environmental factors that contribute unwanted sources of score variation (bias) to clinical performance evaluations.

Summary: Raters have a 1 or 2-dimensional conception of clinical performance and do not recall details. Good news is reported more quickly and fully than bad news, leading to overly generous performance evaluations. Training has little impact on accuracy and reproducibility of clinical performance ratings.

Conclusions: Clinical performance evaluation systems should assure broad, systematic sampling of clinical situations; keep rating instruments short; encourage immediate feedback for teaching and learning purposes; encourage maintenance of written performance notes to support delayed clinical performance ratings; give raters feedback about their ratings; supplement formal with unobtrusive observation; make promotion decisions via group review; supplement traditional observation with other clinical skills measures (e.g., Objective Structured Clinical Examination); encourage rating of specific performances rather than global ratings; and establish the meaning of ratings in the manner used to set normal limits for clinical diagnostic investigations.

Global ratings based on observing convenience samples of clinical performance in vivo form the primary basis for appraising the clinical skills of medical students, residents, and practicing physicians (hereafter called practitioners). Global ratings are subject to several sources of bias that have not been the focus of earlier research and reviews because they involve cognitive, social, and environmental factors attendant to the ratings rather than measurement problems such as instrument design. Research shows that features of measurement instruments account for no more than 8% of the variance in performance ratings.

This review article aims to explore the cognitive, social, and environmental factors that contribute unwanted sources of score variation (bias) to ratings of practitioners. Such bias compromises the value of performance ratings and constrains inferences about the clinical competence of practitioners. To start this we
set the stage by defining clinical competence and describing characteristics that are pertinent to this review. Later, we describe characteristics of the natural clinical environment that lead to unwanted variation in clinical performance ratings.

Next, we establish four quality criteria that clinical ratings must meet to allow strong inferences about medical clinical competence. We also summarize psychometric data from the performance assessment literature in medicine, business, and psychology that document how performance ratings achieve these four criteria and, in the process, better define these criteria.

We then summarize the evidence about cognitive, social, and environmental factors that can bias evaluator ability to make valid and useful inferences about competence from performance ratings. The goal is to identify and describe factors worthy of attention because they are based on well-designed and conducted research studies and, in most cases, have stood the test of replication. It is not possible to describe all studies and findings in detail. Interested readers will have the references to allow further exploration as desired.

Finally, we present a set of 16 recommendations for clinical performance assessment practice and for research based on our analysis of performance rating problems and our review of the literature.

The article is not a review using a single, standard, and highly specific set of methodological inclusion criteria, nor is it an exhaustive review of the performance assessment research literature. Instead, it focuses selectively on cognitive, social, and environmental sources of error that compromise the value of clinical performance ratings as typically collected and the inferences about clinical competence that are based on this information. We try to be comprehensive in reviewing research studies that address these sources of error and to summarize the results when the research design of the studies and the preponderance of evidence earned our confidence. We cite a combination of field studies that document the extent of the problem, and experimental studies (randomized controlled trials) that allow us to disentangle the elements of the problem and document their relative strength. Using the taxonomy of literature reviews proposed by Cooper, the focus of this review is research findings; the goals are to formulate general statements from multiple studies, to bridge the gap in research literature from disciplines involved in performance assessment research, and to identify and refine key issues central to this field of practice and research; the perspective is to provide a neutral representation of the findings; the review describes a purposeful sample of the articles that represent the extensive literature we reviewed; the organization of the review is around a conceptual framework that we developed to guide future practice and research; and the intended audiences are those individuals responsible for the practice of clinical performance assessment and those who do research in this area. Hopefully, the review will also help those practitioners and researchers who focus on performance assessment outside the field of medicine. All factors discussed and recommendations made are based on findings from sound research studies that have also stood the test of replication, taking advantage of meta-analyses when available.

Clinical Competence Defined

Michael Kane defined clinical competence as the degree to which an individual can use the knowledge, skills, and judgment associated with the profession to perform effectively in the domain of possible encounters defining the scope of professional practice (p. 167). This definition highlights a key point about clinical competence appraisal. Evaluators are interested not only in how the clinician performs in observed situations (i.e., judging clinical performance) but also as grounds for generalizing about the practitioner’s ability to perform a variety of other tasks in a range of similar clinical situations (i.e., estimating clinical competence).

Estimating a practitioner’s professional competence requires at least three inferences: (a) judging the quality of the observed performance, (b) drawing general conclusions about the person’s ability to perform in a universe of similar situations based on the observer’s judgment of the observed performance (Kane’s example: drawing general conclusions about skill in delivering babies from the observation of one or two deliveries), and (c) extrapolating from the appraisal context to expected performance in practice situations where the practitioner is not being observed. This third distinction, between what a practitioner can do and what the practitioner typically will do in practice, introduces the contribution of practitioner personal qualities (e.g., dependability, responsibility, ethical beliefs) to the mix of factors (knowledge, skills, and professional judgment) that shape practice behavior.

Clinical Competence Characteristics

Multidimensionality

The term clinical competence leads one to think in terms of a unitary performance capability. This oversimplifies medical practice. Most patient encounters require the practitioner to integrate and perform at least 10 separate component capabilities. These include (a) discerning what information is needed to clarify a patient’s problem; (b) collecting data about the patient (history taking, physical examination, diagnostic tests, procedures); (c) organizing, integrating, and interpret-
ing patient data to derive working diagnoses and refining them as more information becomes available; (d) establishing management goals and evaluating the effectiveness of treatment, adjusting management as needed; (e) performing procedures required for patient management; (f) communicating and relating effectively with patients and their families; (g) working effectively with other health professionals under the range of conditions that make up clinical practice; (h) critically appraising new information for adequacy, quality, and veracity; (i) monitoring personal knowledge and skills and updating these in an efficient and effective manner as needed; and (j) performing consistently in a dependable and responsible manner.

Case Specificity

Research on clinical problem solving,8 clinical performance assessment (primarily using standardized patient technology),9–12 and clinical practice has shown consistently that clinical performance is case specific. Practitioners who excel at diagnosing and managing one clinical problem are not necessarily skillful at diagnosing even closely related clinical situations.13 Practitioners differ in effectiveness in different clinical situations due to differences in experience, training, professional interest, and personal disposition.

Van der Vleuten and Swanson9 summarized the findings from 29 reports of large-scale standardized patient-based examinations of clinical performance conducted on residents and medical students in the 1980s. They concluded that the major source of score variation is due to variation in examinee performance from situation to situation (case to case), and that 16 hr of performance observation across a variety of clinical situations is necessary to achieve reproducible estimates of clinical competence. The studies that formed the basis for van der Vleuten and Swanson's9 conclusions involved (a) carefully selected situations and tasks where observation was tightly controlled; (b) instruments and rater training that assured commonality in rating focus; and (c) ratings that occurred immediately after the performance, limiting forgetting and memory distortion. Because this estimate was made under nearly ideal conditions, it is likely that more hours of performance observation would be required to achieve reproducible estimates of clinical competence in less controlled clinical settings. The van der Vleuten and Swanson9 review suggests that appraisals designed to determine whether clinical performance is acceptable rather than just determine how practitioners rank relative to each other will require observation of a broader clinical sample because the difficulty of the observed encounters must be taken into account.

Field research14,15 involving ratings of physician performance also suggests case specificity. Carline et al.15 recruited 269 interns from six Western states to participate in a study designed to validate the American Board of Internal Medicine (ABIM) certification process. Six professional colleagues (two internists, two surgeons, one head nurse, and one medical director or chief of staff) nominated by each physician were asked to evaluate the performance of that physician in eight areas of clinical competence and to provide a rating of overall clinical skills. The return rate averaged approximately 78% for the raters. Results of this study suggested that 10 to 17 ratings by peers are necessary to yield reproducible global ratings of competence in each competency area. Kreiter et al.16 did a similar study to investigate the measurement characteristics of ratings of medical student clerkship performance using standard clinical evaluation forms. Again, their results documented that ratings depended heavily on the individual rater to whom the student was assigned and on the nature of the patient–student encounter observed. They used decision study results to estimate the reliability of ratings for panels of up to eight raters per student. Panels of eight raters resulted in reliability estimates of at best 0.65. More than eight raters would be required to achieve reliabilities of 0.80, but the article does not allow us to be more precise. Clinical rating research is clear that many observations are needed, and attention needs to be directed toward assuring systematic sampling across the range of normal clinical encounters and tasks.

Context

The typical performance assessment context is an uncertain, uncontrolled, and “noisy” information environment. The most common performance assessment approach is to observe, rate, and judge a practitioner’s clinical performance in a natural clinical setting. The practitioner is observed performing routine clinical tasks under real clinical conditions. A global rating scale containing relatively nonspecific items designed to be used in a range of clinical situations is employed to direct the observer’s attention to common, important aspects of clinical performance and to calibrate the ratings of performance quality. The result of this general approach is that practitioners are observed working up cases that vary in difficulty and in required skill. Likewise, performance standards used for the case may vary because only one expert observes the doctor–patient encounter, and the expert rarely studies the case in depth. Consequently, this method provides a weak basis for comparing clinicians because the practitioners perform different tasks, and the standards used by raters are idiosyncratic.

Contextual noise contributes error (unreliability) to performance measurements. The rater usually has
many responsibilities beyond supervision, training, and evaluation of practitioners. As a result, the rater’s database about a medical practitioner’s clinical performance is often fragmentary and restricted to a small, unsystematic sample of clinical situations and tasks due to limited direct personal contact. Given what has been written about the multidimensionality and case specificity of clinical competence, problems produced by using small, unsystematic, clinical performance samples are magnified. In addition, performance expectations are rarely articulated for the rater or the practitioner. A lack of shared understanding of responsibilities may influence clinical performance ratings. Also, clinical care in hospital and outpatient settings is usually a team responsibility. Therefore, it is often difficult to isolate and appraise the contributions and clinical competence of individual practitioners.

Quality of Clinical Ratings

Observational data must fulfill four quality criteria. First, the observational instrument and process must be accurate. That is, observations must yield scores that are defined by the behavior being observed and recorded, nothing else. Second, the observers must agree. A rating instrument may be capable of accurate use, but rater agreement depends on how it is actually used by observers in a field setting. If only one rater is using the instrument, intrarater reliability is the only issue; but this is rarely the case. When multiple raters use the instrument, interest also resides in agreement among different raters when observing the same performance. The third criterion concerns the ability to generalize from a set of observations to a range of pertinent professional performance situations. As described in the discussion of Kane’s definition of professional competence, the goal of measurement and evaluation of clinical performance is to formulate accurate inferences about clinical competence generally from a series of observations of particular instances of performance. The third criterion addresses this issue. It asks, “To what extent do observations of this set of performances predict how that individual will perform in other situations?”

In addition to accuracy, rater agreement, and the ability to generalize across situations, the focus must include the fourth criterion: utility. Utility addresses the question about whether the generalizations are useful. Does knowing the rating score have any value in predicting future or concurrent performance of interest? Does knowing the score aid in making progress, teaching and learning, licensure, or certification decisions? If not, the ratings have no real value. Utility data also provide the basis for setting meaningful performance standards.

Accuracy

In a randomized, controlled trial designed to determine the accuracy of faculty evaluations of residents’ clinical skills, Noel and colleagues17 had 203 faculty internists watch a videotape of one of two residents performing a new patient work-up on one of two patients. The faculty internists then immediately rated the clinical skills of the resident. One half of the raters used an open-ended evaluation form, and the other half used a structured form that prompted detailed observations of selected clinical skills. The unique contribution of this study was that it investigated and compared attending ratings of the same clinical performance. When observations were not specifically directed through use of a structured form, a condition typical of traditional clinical performance rating practice, participants recorded only 30% of known resident strengths and weaknesses. Strengths were omitted more frequently than weaknesses, an outcome consistent with findings from the general performance assessment literature.18,19 Rater accuracy increased to 60% or greater when participants used structured forms designed to direct attention to specific dimensions of performance. Comparable results were obtained in an earlier, smaller, randomized controlled trial.20

The only other study that addressed accuracy in a direct experimental manner was a study by Pangaro and associates21 done to document the accuracy of standardized patient-produced checklist scores. In this study, internal medicine residents were trained to act as ordinary candidates and to achieve target scores by performing to a predetermined “strong” or “weak” level on checklist items used by raters to record interviewing, physical examination, and communication skills. The strong standardized examinees (SEs) were trained to score 80% correct, and the weak SEs were trained to score 40% correct. Seven SEs took the Standardized Patient (SP) examination along with regular candidates. The mean score recorded by the raters was 77% for the strong SEs and 44% for the weak SEs. A videotape review revealed that most scoring discrepancies were in the area of communication skills.

Achieving the increases in accuracy reported by Noel and colleagues17 and by Herbers and colleagues20 would require standardized patient encounters or, at least, use of rating scales specific to the case being observed and appraised (i.e., rating scales specific to the chief complaint of patients). Obviously, this would come at added cost, extra effort, and added logistical complexity.

There is usually a time lag between observation of clinical behavior and rating of that behavior in the real world of performance evaluation. Delay introduces error into ratings. We know of no controlled study of the effects of time lag on accuracy of clinical ratings, but
laboratory studies reported in the business literature and a survey of psychologists outlined the nature and magnitude of the relation.

In an experimental study with 180 participants, Heneman\textsuperscript{22} demonstrated that ratings recorded 1 to 3 weeks after observation were less accurate than ratings recorded immediately. However, the decrease in accuracy was not large. The delay accounted for only 2\% of the variance in rating accuracy.

A related but “soft” source of evidence confirms the importance of immediate recording to maximize rating accuracy. Kassin et al.\textsuperscript{23} conducted a survey of 64 psychologists who are frequently asked to testify about the reliability of eyewitness testimony. The survey asked the psychologists to list findings that are adequately documented to present in court. The expert psychologists agreed there is sufficient evidence to support the conclusion that event memory loss is greatest right after the event and then levels off with time.

**Rater Agreement (Intra- and Interrater Agreement)**

Intrarater agreement is based on having a single rater evaluate the same (e.g., videotaped) performance on at least two different occasions. Differences in ratings reflect changes in rater attention, perspective, standards, or mood. These studies provide a good estimate of how much variability is introduced into a performance measurement as a result of these factors.

Kalet et al.\textsuperscript{24} had faculty raters rerate videotapes of 21 medical student interviews after a period of 3 to 6 months. They found substantial agreement between the earlier and later ratings only when raters judged the information obtained (i.e., data collection) during the interview. Even here, the level of agreement was far from perfect. There was little to no agreement in the earlier and later ratings of the interviewing process.

Interrater agreement is based on having different raters evaluate the same performance. Under these circumstances, differences in ratings reflect differences in rater focus and rigor as well as the sources of variability addressed in intrarater reliability studies. Weekley and Gier\textsuperscript{25} examined ratings gathered from world-class figure skating judges during the 1984 Olympics. These results provide some indication of the theoretical upper limit on intrarater agreement using human judges. Olympic judges have at least 15 years of judging experience. Criteria for judging are clearly spelled out, and judges are periodically trained and retrained to use them. Nine judges are employed to observe and appraise the same performance. Olympic judges have no conflicting responsibilities during judging, and they receive continuous feedback about the consistency of their ratings with those of other judges because all ratings are posted immediately. This serves as a form of feedback that can refine agreement among future ratings and also is likely to increase the motivation to provide accurate ratings.

Interrater reliability coefficients were consistently high, ranging from 0.93 to 0.97 for the full complement of judges. More detailed analyses demonstrated that the judges were able to agree in their evaluation of the skaters, but primarily in a global sense. Judges’ ratings of skaters did not differ much for the two dimensions rated, technical merit and artistic impression. The rating differences between technical merit and artistic impression were larger in the short program where skaters were required to perform prescribed maneuvers, reinforcing the view that careful definition of tasks performed and rated helps raters to differentiate among performance dimensions. These results suggest that the ceiling on rating reliability using human judges is high. Virtually all of the variability in ratings was attributable to differences in skating performance.

Elliot and Hickam\textsuperscript{26} had groups of three physician faculty members observe the same student while the student performed a physical examination on a patient. As in the Weekley and Gier\textsuperscript{25} study, observers were familiar with the task, had limited environmental interruptions to disrupt the observation, and used a checklist with careful description of the tasks and finite classification categories for rating performance. Elliot and Hickam\textsuperscript{26} found that observers did not agree when judging 17 of the 53 physical examination maneuvers performed (32\%). There was good interrater agreement for 18 of the remaining maneuvers and moderate agreement for the other 18. Agreement was lower on items requiring palpation and percussion, characteristics that are not subject to complete evaluation by observation alone. Elliot and Hickam\textsuperscript{26} also reported that trained nonphysician observers were able to reliably evaluate 83\% of the skills that were reliably assessed by faculty.

Kalet et al.\textsuperscript{24} also had four judges each rate medical interviews. They concluded that interrater agreement was poor when judging the overall interview as well as when judging the information obtained and the interviewing process.

In the Noel et al.\textsuperscript{17} study described earlier, medical faculty raters were also asked to appraise the overall clinical performance of the resident. Raters disagreed about the residents’ overall clinical performance. For the first resident, 5\% of the judges rated the overall clinical performance as unsatisfactory, whereas 5\% rated the performance as superior. Twenty-six percent rated the performance as marginal, and 64\% rated the performance as satisfactory. For the second resident,
the ratings were 6% unsatisfactory, 42% marginal, 49% satisfactory, and 3% superior. Agreement in overall ratings was not improved among raters who filled out the structured form. These results strongly suggest that judges use different criteria and standards to judge clinical performances. Therefore, multiple ratings need to be collected and combined to assure a fair and representative overall appraisal of clinical performance. These results are a conservative estimate of actual clinical performance assessment problems because the raters in this study all observed the entire performance and were able to complete their appraisals immediately under circumstances where no competing responsibilities interfered. Again, parallel findings were observed in the Herbers et al. study.

Martin and colleagues compared the score established by individual physician raters with a gold standard based on the average score of three physician raters who rated the same encounters from a videotaped record. This allowed the investigators to determine the absolute score deviation from the gold standard as well as conduct correlational analyses that compare relative rankings of individuals. The average correlation of the individual physician raters with the gold standard was approximately 0.54 with a fair amount of variation among clinical situations. The unique contribution of this study was to document that individual physician scores were almost always overestimates of true clinical competence (estimated by the gold standard consensus ratings). This tendency to give medical practitioners the benefit of the doubt has also been established in a study by Vu and colleagues.28

MacRae and colleagues compared physician ratings of 120 videotaped medical student clinical encounters across four cases using an eight-item rating scale covering information acquired and interviewing process. They found that the interrater reliability coefficients ranged from 0.65 to 0.93 with an average reliability of 0.85. These results demonstrate that good interrater agreement is possible when judging clinical performance. One factor that probably contributed to the high level of agreement is that the raters collaborated in developing the rating scales used in this study, resulting in a shared definition of good clinical performance.

The differences in overall appraisal of the same performance documented in this section are most likely due to the combination of two factors. First, different raters focus on (differentially weight) different aspects of clinical performance. Second, different raters have different expectations about levels of acceptable clinical performance. Anything that reduces these differences, such as in the MacRae et al. study, tends to improve interrater agreement.

**Accuracy in Estimating Competence (Estimating From a Discrete Set of Observations to Performance Across a Range of Tasks and Situations)**

Interrater agreement studies using ratings by judges who have observed different practitioner performances combine variance due to differences in judges’ expectations with variance due to differences in demands (task and case characteristics) of the observed sample of behavior. Viswesvaran et al. used meta-analytic methods to investigate interrater and intrarater reliabilities of job performance ratings reported in 15 primary organizational psychology and business journals from their inception through January 1994. This resulted in studying 215 research reports. Viswesvaran et al. established that the mean interrater reliability estimate for ratings by supervisors equaled 0.52, with the majority of interrater reliability coefficients falling between 0.50 and 0.54. They reported that the mean intrarater reliability coefficient was 0.81, with most coefficients ranging from 0.78 to 0.84. Based on the differences between inter- and intrarater reliability coefficients, Viswesvaran and associates concluded that approximately 29% of the variance in ratings is attributable to differences in the perspective of different raters, including the sample of behavior they observed; whereas 5% is due to variations in a single rater attributable to changes in moods, frames of reference, or other variables.

Recently, authors of a number of studies on clinical performance evaluation have been influenced by generalizability theory and analytic methods. This work has established the number of ratings needed to achieve a generalizability coefficient of 0.80 that is accepted as the minimum desired level for use in making progress decisions. Carline et al. collected 3,557 ratings of 328 medical students who had taken an internal medicine clerkship. They concluded that 7 independent ratings would be needed to judge overall clinical performance. The amount of rater experience did not change the results. Similar results have been reported by Kreiter and colleagues (more than 8 ratings needed), Kwolek et al. (7–8 ratings), Ramsey et al. (11 ratings), Violato et al. (10 ratings), and Kroboth et al. (6–10 ratings). All reports agree that somewhere between 7 and 11 ratings are necessary to achieve a generalizable global estimate of competence when raters are basing ratings on a nonsystematic sample of observations.

Although there has been only limited research on the component skills included under the broader category of clinical competence, it is reasonable to expect that these abilities will develop at different rates and may differ in their stability across situations. The work that has been done suggests that different numbers of
observations will be required to establish stable estimates of clinical competence in various clinical competence areas. Carline et al. established that the number of ratings required to achieve stable estimates of competence in each of the eight areas reflected in their items ranged from 10 to 32 ratings, with an average of 18. The areas where the least agreement occurred included ratings of (a) verbal presentations; (b) written communications; and (c) management of patients with simple, uncomplicated problems (probably because these are managed in the office where care is less subject to observation than in the hospital). Because this investigation was a field study, it is not possible to determine whether these findings are due to less stable practitioner performance or to rater variability associated with lack of rater opportunity to observe. Carline et al. also reported that the number of independent ratings needed to reliably judge various individual aspects of clinical performance varied dramatically. To illustrate the range of requirements, these investigators reported that 7 ratings would be needed to judge data gathering reliably, whereas 27 would be needed to rate interpersonal relationships with patients. Wenrich et al. observed similar results when studying nurses’ ratings of internists. They established that somewhere between 11 and 24 nurse ratings would be necessary to establish a generalizable estimate (generalizability coefficient = 0.80) of clinical competence depending on the competence dimension being investigated. Interpersonal and professional behavior characteristics required more ratings than did cognitive and technical attributes. Meta-analytic reviews of the general performance assessment literature produced similar findings. More ratings are needed to provide stable estimates of communication and interpersonal competence than for specific skills (e.g., clinical skills).

We believe that the discrepancies in number of ratings needed for interpersonal and professional behavior and cognitive and technical attributes are best explained by the conceptual framework offered by Ginsburg and associates. These investigators noted that we tend to talk about people as being honest or dishonest, professional or unprofessional rather than talk about their behaviors. In other words, professionalism is thought of as a set of stable traits. They argued instead that professional behavior is much more context dependent than has usually been acknowledged. The implication of this is that generalizations about professional behavior require sampling much larger numbers of events (contexts and behaviors) than is true for some other aspects of competence. More research regarding the nature of aspects of clinical competence will help to increase our understanding of these individual competencies and will allow further refinement of performance assessment practice.

It is clear that multiple observations of students, residents, and physicians are needed to reliably estimate clinical competence. The results in the studies reviewed here suggest that a minimum of 7 to 11 observer ratings will be required to estimate overall clinical competence. More ratings will be necessary to obtain stable estimates of competence in more specific competence domains (e.g., data collection, communication, responsible professional behavior), with the most ratings (up to 30) needed in the noncognitive areas.

A few studies have been reported that investigate using nurses and patients to appraise aspects of clinical competence. We have included these studies under the heading of Accuracy in Estimating Competence because we believe that these ratings broaden the dimensions of competence observed more than they change the characteristics of raters. Wenrich and associates compared ratings of physicians by nurses and physician peers. They found that the correlation among these ratings for individual clinical performance components ranged from 0.46 to 0.72, with the average correlation being 0.59. Nurses’ ratings were generally lower than those of physician peer raters for humanistic qualities, and their ratings were higher for medical knowledge and verbal communication. These results suggest that nurses’ ratings are not a replacement for physician ratings. However, nurses ratings do expand the number of perspectives incorporated into the estimates of clinical competence and broaden the range of clinical situations and competencies observed and judged. Nurses also are more likely to have the opportunity to observe how practitioners typically perform.

Tamblyn and associates investigated the feasibility and value of using satisfaction ratings obtained from patients to evaluate the doctor–patient relationship skills of medical residents. They found that patient satisfaction ratings provide valuable information about a resident’s ability to establish an effective doctor–patient relationship. However, these investigators noted that approximately 30 patient ratings would be necessary to obtain a generalizable estimate of resident skill in these areas (0.80 generalizability coefficient), although a smaller number of ratings would be enough to detect outliers. The larger number of required ratings speaks to the increased heterogeneity of patient raters and raises feasibility issues. However, the use of patient satisfaction ratings provides the perspective of the client. Church’s findings from a business environment suggested that service providers and their clients have different perceptual frames of reference than internal observers. In this regard, the addition of information from patients should provide an important perspective that has not often been captured systematically.

Utility

The bottom line for any measurement instrument revolves around its usefulness. A useful clinical perfor-
CLINICAL PERFORMANCE RATINGS

Hojat et al. \textsuperscript{40} investigated how effectively medical student clerkship ratings predicted residency supervisors’ ratings of clinical performance. The investigators sought to determine if clerkship ratings provided information that was helpful in predicting which students performed poorly or well during their first year of residency. They found that 42\% of students who received clerkship ratings that put them in the bottom 25\% on global ratings of clinical competence were also in the bottom fourth based on residency program ratings of data gathering and processing skills. Only 17\% of those in the bottom fourth on medical school clerkship ratings were found to achieve residency ratings that put them in the top fourth in their residency program. Thirty-seven percent of those individuals who received clerkship ratings that put them in the top performance quartile were found to achieve residency ratings that placed them in the top quartile. Conversely, only 14\% of those individuals ended with residency ratings that placed them in the bottom fourth. Put another way, the 1,704 individuals were three times more likely to be classified in the same quartile during residency training and medical school as to be classified at the other extreme (i.e., classified in the bottom fourth during residency when they had been classified in the top fourth during medical school).

Data from rating instruments acquire meaning by documenting consistent relations between scores on the instrument and scores from other measures of interest. Given this information, it is possible to use scores to make progress decisions with confidence that the scores have implications for concurrent or future performance states. In the case of the Hojat et al. \textsuperscript{40} research, it becomes possible to anticipate subsequent clinical performance during residency training. This information is critical for making informed decisions about medical student progress. There are surprisingly few examples of this type of investigation in the literature that we studied.

One of the primary obstacles to research on clinical ratings is the lack of a readily available, consensual gold standard for judging clinical competence. One approach used in standardized patient research is to develop a patient-based problem with known findings and have a group of experts decide in advance what information should be elicited to diagnose the case effectively, what diagnosis is most likely, and what management is indicated. The case can then be used to test the clinical performance of practitioners. Recent research using patient simulators has demonstrated the feasibility, utility, and validity of such clinical evaluations. \textsuperscript{41–47} A second approach to the gold standard problem is evaluation of a single performance by a group of expert judges who all observe the same performance and then rate the performance independently (perhaps using videotape). A third, more common approach in traditional clinical performance assessment involves a comparison of the ratings of a group of expert judges who observed the same practitioner across an uncontrolled and undocumented range of situations and tasks in natural environments. Although this approach has most of the weaknesses described earlier, it is the most prevalent. Cronbach \textsuperscript{48} contributed to refining this approach conceptually by proposing that rater performance be analyzed to determine whether (a) an individual rater tends to rate too high or too low averaged across all raters and items (leniency–stringency), (b) the rater can accurately distinguish among practitioners based on the candidate’s total job performance (i.e., the rater can accurately rank practitioners relative to each other), (c) the rater can accurately determine whether a group of practitioners are better at one aspect of their jobs or another, and (d) the rater can diagnose accurately the strengths and weaknesses of individual practitioners. This framework for thinking about and evaluating the quality of ratings is appealing and practical. Where called for, it is used in subsequent sections of this article.

Cognitive, Social, and Environmental Sources of Bias in Clinical Ratings

Research in experimental psychology has investigated the nature of human memory since the late 19th century. Since the 1980s, research in performance assessment has taken conceptual cues from this research. A number of experimental studies have investigated how humans store information and use that information in performance assessment settings. The research results provide partial answers to the following two questions. How are rating data organized? What are the mechanisms of data recall?

Data Organization

Research has been conducted to determine the strategy raters use in organizing and storing the information they acquire and to determine which organizational strategies result in the most accurate recall and ratings.

These research results remind us of the limits of human memory and the implications for performance assessment. Nathan and Lord \textsuperscript{49} concluded, “Even under optimal conditions many subjects will not or cannot independently use five dimensions to rate performance” (p.112). Evidence supporting this conclusion was present in the Weekley and Gier \textsuperscript{25} study of world-class figure skating judges during the 1984 Olympics. Their
analyses demonstrated that the judges agreed about their evaluations of the skaters, but chiefly in a global sense.

Cafferty et al.\textsuperscript{50} reported that raters organized information either by persons (performances of one person on all pertinent tasks, followed by performances for a second person) or by tasks (performances of all persons for one task, followed by performances of all persons for a second task, etc.). Not surprising, acquisition by person led to information being organized in memory by persons, whereas information acquired by task was organized in memory by tasks. Most important, raters who did not organize information in memory or were inconsistent in their organizational strategy had the least accurate recall and ratings. DeNisi and Peters\textsuperscript{51} investigated the effects of structured diary keeping on recall of performance information and on performance ratings. They found that raters who kept diaries structured by persons or by performance dimensions were better able to recall performance information than were those who kept no diary even when they were not allowed to refer to the diary. Raters who kept diaries provided (a) lower overall ratings, (b) ratings of individual practitioners that varied more from dimension to dimension (provided a more varied profile for each person), and (c) ratings that discriminated more among ratees. Raters who organized their diaries by persons believed that their ratings were more accurate and defensible than did raters in the other conditions, and they believed that the use of diaries helped them in the appraisal process. Raters who were required to maintain diaries, but where no structure was imposed consistently, organized their diaries by persons.

Research by Williams and colleagues,\textsuperscript{52,53} in a laboratory setting, suggested that raters who acquired and stored performance information in an unorganized fashion can be prompted to structure the information stored in memory before a rating task by having them recall pertinent performance incidents and record them according to a given format. This structuring resulted in recall and rating accuracy results similar to those obtained when the information was initially obtained in a structured manner. DeNisi and Peters\textsuperscript{51} found that raters in a free recall situation chose to organize the performance information by persons. Most important, they found that raters asked to organize the information before the rating task provided less elevated and more discriminating ratings and reported more positive reactions to the rating task. This suggests that raters can be prompted to structure performance information in memory just before the appraisal task and realize a rating benefit without adding the logistic demands of a diary-keeping requirement. Although we assume a posture of healthy skepticism about these findings until they stand the test of multiple replications, they are interesting and promising.

Therefore, a series of laboratory and field studies suggest that (a) organization of information in memory is important for recall and rating accuracy; (b) recall and rating accuracy can be improved by prompting the rater to organize information before the rating even if the information was acquired in an unorganized fashion; and (c) raters prefer to store information organized by persons, and this strategy produces the most accurate recall and ratings.

Data Recall

Stored information about practitioner clinical performance can be recalled and used by raters to formulate ratings in two ways. First, recall may involve specific details about practitioner performance in specific clinical situations. Second, recall may be based on observations of specific events, forming a general impression of the effectiveness of performance in that situation, and storing only the general impression. In the general case, much of the specific information used in forming an impression is lost. When asked to make an appraisal, the rater recalls the pertinent general impression data and uses it for making a judgment. Evidence regarding how raters perceive and recall clinical performance data is instructive. Results of factor analytic and correlational studies suggest that raters provide global ratings of clinical competence rather than differentiate individual aspects of competence. Verhulst et al.\textsuperscript{54} found that supervisor’s ratings (13-item rating form) of first-year resident performance revealed that raters differentiated and rated two aspects of clinical performance: clinical skills and professional behavior. These two factors together accounted for 83.6% of the common variance in ratings, with the clinical skills factor accounting for 49.4% and the professional behavior factor accounting for 34.2% of the common variance. The results were virtually identical each of the 7 years of the study.

In their study of peer ratings, Ramsey and colleagues\textsuperscript{32} also found that raters have a two-dimensional conception of clinical skills. One factor represented cognitive and clinical management skills and the other represented humanistic qualities and management of psychosocial aspects of illness.

Other studies reported similar results suggesting that clinical performance ratings reflect a general impression of clinical performance rather than a differentiated view of individual competencies.\textsuperscript{33,35,54–58} These results suggest that either the individual competencies that make up clinical competence are highly correlated (i.e., individuals who perform one competency well are likely to perform well in the other competency areas); or that individual competencies are not highly correlated, but raters perceive them as such and rate them accordingly. Nathan and Lord\textsuperscript{49} summarized their research on performance assessment by stating:
The true correlation among dimensions related to any naturally occurring category will usually be nonzero, because categories are built from natural clusters of characteristics. On the other hand, the correlation among rated dimensions will almost always be higher than true correlations, due to the nonrandom errors or simplifications of raters. (p.112)

This is an area where a number of large-scale studies have been done regarding clinical performance assessment. Results from these studies show consistently that raters have either a one- or two-dimensional conception of clinical skills and do not differentiate further among the specific dimensions of clinical performance using current instruments and procedures. Efforts to elicit more differentiated ratings may result in rater time and effort being wasted, and this may result in inaccurate characterizations of true performance.

**Frequency of Rater Observation**

In a survey of 336 internal medicine residents from 14 New England residency programs, Stillman and colleagues found that 19% of the residents reported that no faculty member ever observed them performing a complete history and physical examination during medical school, and 30% reported they had never been observed performing a complete history and physical in residency. Over one half reported that they had been observed less than three times during medical school; and 45% reported that they had been observed once during residency, presumably during the required ABIM Clinical Evaluation Exercise (CEX). Burdick and Schoffstall reported similar results in a survey of 514 residents in 21 emergency medicine residency programs throughout the United States.

These results suggest that the frequency of direct clinical performance observation at the medical school and the residency level is very low. To the best of our knowledge, no one has reported on the total number of hours of observation per student or resident underlying clinical performance evaluation. Based on surveys of 3,600 business organizations throughout the United States, Bretz et al. reported that business supervisors commonly spend about 7 hr per year assessing the performance of each employee. This includes any special observation, other data collection, filling out assessment forms, and communicating with each employee. Because middle- and upper-level human resource managers filled out the survey, one might expect that these results paint a favorable picture of the process. We know of no comparable survey of clerkship directors, residency program directors, or health maintenance organization directors, but we have no reason to expect that more time is spent in clinical performance assessment. Therefore, this section is directed toward investigating available evidence about the impact of frequency of observation on the accuracy and reproducibility of ratings.

In a study of almost 10,000 pairs of raters from 79 business organizations, Rothstein established that interrater agreement increases as the length of exposure to the ratee increases. Interrater agreement improved slowly and reached only about 0.60 with 10 to 12 years of opportunity to observe long-time employees. Heneman, in a randomized controlled trial, also found that longer observation periods yield more accurate ratings. Finally, eyewitness testimony experts agreed that there was sufficient evidence to testify in court that the less time an eyewitness has to observe an event the less accurate is the memory of the event. Therefore, there is evidence from a range of situations to reinforce the idea that the frequency of observation influences the reproducibility and accuracy of ratings. The quality and utility of rating data is likely reduced at the low levels of observation that are common in clinical performance assessment.

**Focus of Attention**

One constructive role that a clinical performance assessment instrument can play is to direct the attention of raters toward particular dimensions of competence. This may increase the level of rater agreement and the accuracy of ratings.

In the randomized controlled trials described earlier, Noel and colleagues and Herbers and colleagues found that participants recorded only 30% of resident strengths and weaknesses using global rating forms. Strengths were omitted more frequently than weaknesses, a finding consistent with results from the general performance assessment literature. Rater accuracy increased to 60% or greater when participants used structured forms designed to focus attention on specific performance dimensions. Noel and colleagues reported the percentage of raters who detected each strength and weakness but the number of items is too small to allow generalizations about the attributes of performance that raters observed. More work of this type is needed to determine what expert raters attend to when rating clinical performance. If, as we suspect, the scope is limited to a few competencies (e.g., clinical skills, professional behavior) then efforts need to be directed to expanding rater focus.

Woehr and Day found that when raters are told about task rating dimensions in advance they perform better in rating that performance. Day found that raters were better able to distinguish between effective and ineffective behaviors when they had advance access to the rating instrument. Specifically, raters were able to more accurately identify behaviors that did not occur. Raters without prior knowledge of task rating dimensions were more dependent on a general impression of effectiveness.
Rater Reporting

Judgments are private evaluations of a practitioner's performance, whereas ratings are a public record. For a variety of reasons the two may not always coincide.

In a review chapter, Tesser and Rosen64 described a bias among communicators to transmit messages that are pleasant for recipients to receive and to avoid transmitting unpleasant messages. They labeled the tendency to avoid transmitting bad news the “Mum Effect.” The review recounts the results of studies conducted by themselves and other investigators to probe this bias directly. Tesser and Rosen64 concluded that good news is communicated more frequently, more quickly, and more fully than bad news. They also reported this is true across a wide range of situations whether the communication is in person or in a less direct manner (e.g., written communication). This Mum Effect would seem to have clear implications for supervisors’ communications of poor performance to practitioners.

A number of studies have been conducted to determine the impact of the Mum Effect on supervisors’ communications about performance of employees and trainees. In a laboratory study involving performance assessment, Benedict and Levine65 found that participants scheduled low performers for feedback sessions later and evaluated these performers with more positive distortion. Two field studies suggested similar results. Waldman and Thornton66 found that ratings used for promotion and salary increases and those that were to be shared with employees were more lenient than confidential ratings of the same employees. Harris et al.67 had supervisors rate employees to aid in validating a performance examination. The supervisors had earlier rated the same employees for purposes of promotion and salary increases using the same instrument. The investigators reported that slightly more than 1% of the raters were significantly more stringent in their ratings as they apply to clinical performance ratings. Students with true clinical ability at the 50th percentile would receive ratings from these lenient raters that would rank them at the 76th percentile or higher. Conversely, 14% of the raters were classified as significantly more stringent in their rating patterns. Ratings by these stringent raters would rank average students (true clinical ability at the 50th percentile) at the 23rd percentile or lower. However, the total group of raters was normally distributed regarding stringency of ratings. Therefore, where multiple ratings are available for each student, the average rating should approximate the student’s true clinical competence. Further, the investigators noted that the format of rating forms was not related to the tendency to provide extreme ratings. Ryan et al.71 studied emergency medicine supervisors and found that they also showed large variations in both leniency and range restriction, and that differences in individual evaluator scoring leniency were large enough to have a moderate effect on the overall rating received by the resident.

Inflation of Ratings

Based on surveys of 3,587 business and industrial organizations, Bretz et al.69 found that 60% to 70% of an organization’s workforce received ratings in the top two performance levels on the organization’s performance rating scale. We know of no directly comparable large-scale survey regarding medical clinical performance assessment. However, there are a number of studies that report similar outcomes, and the results are consistent across the studies found. Magarian and Mazur72 surveyed all internal medicine clerkship directors in the United States and received completed surveys from 81% of them. Survey results indicated that 75% of students in internal medicine clerkships using an A through F grading system received ratings in the top two performance categories. In clerkships supporting the finding that supervisors are likely to distort ratings of their employees when they believe employees will see the ratings or when the ratings will be used for promotion or salary increase purposes. The Mum Effect research suggests that distortion in communications may occur even when there are no consequences for the employee. In all cases, the tendency is for ratings to be increased. We are not aware of any studies of the Mum Effect or of the degree of distortion in ratings as they apply to clinical performance ratings.
Objective test results generally lower grades.74 Higher if clinical ratings alone were used for grading. Percentages in the top two grading categories would be higher if clinical ratings alone were used for grading. Objective test results generally lower grades.74

Shea et al.75 described program directors’ ratings of all 17,916 internal medicine residents who sought certification in internal medicine during the years 1982 through 1985. They reported that the mean rating was 6.5 on a 9-point scale. Six is a “satisfactory” rating and 7 is a “superior” rating according to the scale used. Residents with “unsatisfactory” program directors’ ratings were excluded from this study. Therefore, the average rating is raised artificially by a small degree. However, it is clear that the ratings assigned are concentrated in the higher scores on the scale.

Ramsey and associates76 reported that 91% of all practicing physicians in their study received a rating of 7 or above on a 9-point scale from peers. The mean clinical performance rating for all physicians was 7.71 with no physician receiving a rating below 6. Other studies16,55,71 reported similar findings across a range of practitioner rating situations (medical students, residents). Carline et al.15 found that out of 6,932 ratings in their study, only 136 or 2.0% were poor or fair ratings. Twenty-two percent of the physicians (58 of 269 physicians rated) received all of the 136 poor or fair ratings, with 15 physicians receiving three or more low ratings.

Some other evidence exists indicating that individual ratings of clinical performance are inflated. We reported earlier that Martin and colleagues27 found that individual physician ratings were almost always overestimates of true clinical competence as established by a consensual gold standard panel. Vu and colleagues28 established this tendency for individual raters to give rates the benefit of the doubt as well. Schwartz et al.77 reported that faculty ward evaluations of fourth-year surgery residents were inflated and did not correlate with other, more objective measures of clinical competence (i.e., Objective Structured Clinical Examination [OSCE] scores).

Speer et al.78 reported on their efforts to modify the clinical rating process in an internal medicine clerkship toward the goal of reducing grade inflation. In the original grading system, the average grade assigned to students by preceptors was an A. The revised system involved using a more behaviorally anchored definition for each grade and by assigning final grades through an “objective group review.” This revised system resulted in students receiving an average grade of B+. The authors documented that preceptors believed the average student should receive a grade of B. Therefore, the revised grading approach was more consistent with preceptor expectations. Although this study does not establish the relative contributions that behavioral anchors and the group review process contributed to decreasing the grades assigned, a study by Hemmer et al.79 supports our belief that the objective group review process was the more important factor. They found that a group review process significantly improved the detection of unprofessional behavior. In almost 25% of the cases, the group review process was the only evaluation method that detected unprofessional behavior.

Metheny56 reported that first-year residents in obstetrics-gynecology were more lenient than other residents in their grading of medical students. These results were based on findings in one clerkship at one medical school. It would be informative to have a larger scale study to determine whether such variables as amount of clinical supervising experience have systematic effects on ratings assigned.

Failure Rates

Another way to study the issues of leniency and stringency and grade inflation is to look at failure rates. The Liaison Committee on Medical Education conducts a survey of all medical schools in the United States each year that normally yields a 100% response rate. The results of the survey80 indicated that attrition from U.S. medical schools continues to be about 1% of total enrollment. Forty-one percent of those students who left medical school during the 1996 to 1997 year were either dismissed for reasons of academic failure or withdrew in poor academic standing. Of these students, only 31% were in the clinical years of training. Therefore, there were 95 students in the clinical years who were dismissed for academic failure or withdrew in poor academic standing from the 125 U.S. medical schools during the 1996 to 1997 academic year. This is less than one medical student per U.S. medical school. Therefore, the likelihood of a student leaving medical school for academic performance reasons during the clinical years of training was very low.

In an Association of American Medical Colleges survey of representatives from 10 medical schools, Tonks and Buchanan81 reported that one of the most commonly reported problems was unwillingness to record negative performance evaluations. A second was unwillingness to act on negative evaluations. Results from the Magarian and Mazur72,73 surveys of internal medicine clerkship directors support the Tonks and Buchanan81 findings. Thirty-six percent of the clerkship directors responding could not recall having ever given a failing grade in their clerkship. The highest failure rate reported for a medical school class was 10% (two schools). Failure rates in other studies were comparable. Yao and Wright82 reported that 6.9% of internal medicine residents were reported to have significant enough problems to require intervention by
someone of authority. They also reported that 94% of residency programs had problem residents. Kwolek et al.\textsuperscript{83} reported that 1.3% of ratings in a surgery clerkship were unsatisfactory or marginal, and 5% of interns in the same program were rated as deficient.\textsuperscript{77} Schueneman et al.\textsuperscript{57} reported that 9 out of 310 (3%) surgery residents were asked to leave their residency training program over a period of 15 years. Vu et al.\textsuperscript{74} studied failure rates in six required clerkships (family medicine, internal medicine, obstetrics–gynecology, pediatrics, psychiatry, and surgery) in one medical school over a period of 5 years. They reported that the average failure rate was less than 1% in family medicine, 1% in obstetrics–gynecology and in psychiatry, 5% in surgery, 7% in internal medicine, and 8% in pediatrics (average failure rate: 3.67%). However, 70% of the failing students received passing marks based on clinical performance ratings alone. In these cases, performance on an objectively scored knowledge examination (the NBME subject test) was the factor that determined their failure. Jain et al.\textsuperscript{84} reported the results of a survey of physical medicine and rehabilitation residency programs. Eighty-three percent of program representatives responded. Only 50% of programs reported rating at least one resident unsatisfactory during a clinical rotation in the 3 years prior to the survey. Only 11% had reported one resident’s overall performance as unsatisfactory to the parent organization.

As mentioned earlier, results from a study in a large business organization\textsuperscript{67} indicated that 1.3% of 223 rates were rated as in need of improvement on regular ratings of employees for promotion and salary increases. When these same employees were rated to validate a performance test, a situation where presumably raters perceived no negative consequences of their ratings for the employees, 6.7% were rated as needing improvement. Therefore, the true rate of cases where performance is inadequate may be closer to 7% than to the 1% rate, which normal ratings yield.

**Rater Calibration and Training**

A common response to the perception that clinical performance ratings are less accurate and reproducible than they should be is to suggest implementing a rater training program. We found very few studies in medical clinical performance evaluation that addressed the effectiveness or cost effectiveness of rater training. Both studies that we found suggested that training may not lead to gains commensurate with the cost of the program. Noel and colleagues\textsuperscript{17} included a training condition in their randomized controlled trial that studied the accuracy of raters in appraising clinical performance. They found that physicians who received training were no more accurate than other raters who participated in the study but without training. Newble et al.\textsuperscript{85} compared the interrater agreement for groups of physicians with no training, moderate training, and more extensive training. They reported that training did not improve the low rate of interrater agreement. However, interrater agreement was improved substantially by identifying the most inconsistent raters and removing them from the analysis. This approach of culling inconsistent raters may be a more cost-effective way of dealing with rater reliability problems.

The limited number of studies that have addressed the issue of rater training for clinical performance evaluation should be judged cautiously. The findings may be the result of the type or quantity of training provided. To illustrate, the training in the Noel et al.\textsuperscript{17} study only involved viewing a videotape. There was no practice and feedback component. The Newble et al.\textsuperscript{85} study involved a ½ hr training session with no feedback for the moderate training group and a 2½ hr training session including practice and feedback for the extensively trained group. The amount and type of training in the Newble et al.\textsuperscript{85} study is more in line with what would be expected to result in a positive training outcome, but it may take even more training to yield the desired improvement in reliability.

Many more studies investigating the effect of training on rater performance have been done in business settings. Woehr and Huffcutt\textsuperscript{86} reported a meta-analysis of 29 rater training studies. These studies yielded 71 meaningful comparisons of training programs with control groups. Most important, Woehr and Huffcutt\textsuperscript{86} sorted out four different types of training programs and investigated their effects separately. The most common programs (28 comparisons) were rater error training programs designed to minimize errors of leniency, central tendency, and halo by making raters aware of these tendencies. These programs proved to be moderately effective in reducing halo error and somewhat less effective with respect to rater leniency. Rater error training programs also yielded a modest increase in rater accuracy with some being substantially more effective than others.

The second type of rater training was performance dimension training, which was designed to train raters to recognize and use appropriate dimensions of performance in their ratings. The idea behind this training is that raters will make dimension-relevant judgments as opposed to more global judgments. Woehr and Huffcutt\textsuperscript{86} reported that performance dimension training was moderately effective at reducing halo error but less effective with respect to increasing accuracy. This training procedure actually led to a slight increase in rater leniency.

**Frame of reference training** programs, the third approach, trains raters with regard to performance standards as well as performance dimensionality. These programs typically provide raters with samples of behavior representing each dimension and each level of performance on the rating scale. The programs also in-
clude practice and feedback. These programs proved to be the single most effective training strategy for increasing accuracy of rating and observation. They also yielded small decreases in halo and leniency.

Finally, behavioral observation training programs focus on improving observation of behavior rather than evaluation of behavior. There have been few (four) such studies reported. Like frame of reference training, behavioral observation training programs were found to produce a medium to large positive effect on both rating and observational accuracy. These studies did not investigate effects on either leniency or halo.

A summary of results from Woehr and Huffcutt suggests that rater training programs have promise as a means of improving performance of raters observing and appraising clinical performance. However, the rater training studies in medical clinical evaluation have not been effective most likely due to their short duration, no practice, and no feedback. This suggests either the training program characteristics have not been optimized or that physician raters are impervious to training.

**Practitioner (Ratee) Characteristics**

This section addresses practitioner characteristics that can affect performance ratings but are unrelated to practitioner performance ability. Practitioner characteristics other than performance ability can be subdivided into two categories: those that are relatively fixed and those that are more transitory and can be manipulated by the practitioner. Authors have used such terms as impression management and social control to describe the overt manipulation of transitory variables for personal gain.

**Fixed Personal Characteristics of Practitioners**

We found no studies investigating the impact of fixed personal characteristics (e.g., ethnicity, age, gender, physical attractiveness) on clinical performance ratings in medicine. However, several studies involving these variables have been done on performance evaluation ratings in the military and in business. Most of the reported research investigates the impact of rater and practitioner ethnic background. A few studies have investigated practitioner age and its impact on assigned ratings. Finally, one study investigated ratee attractiveness and affect.

The most comprehensive and informative studies of ethnicity have been those by Pulakos et al. and by Sackett and Dubois. The Pulakos et al. study involved 39,537 ratings of 8,642 army enlisted personnel in a variety of jobs. They performed separate analyses for the data set where both an African American and a Caucasian rater evaluated ratees. This data set included 4,951 ratees (1,663 African American and 3,288 Caucasian). Sackett and Dubois conducted a similar study using a data set including ratings for 36,000 individuals holding 174 civilian jobs in 2,876 firms. As in the Pulakos et al. study, Sackett and Dubois performed separate analyses for the 617 ratees (286 Caucasian, 331 African American) where both an African American and a Caucasian rater had submitted ratings for a ratee. Sackett and Dubois did a combined analysis that incorporated the data from the Pulakos et al. study and their study. They reported that African American and Caucasian raters provided virtually identical ratings of Caucasian ratees. African American ratees, on the other hand, received ratings from Caucasian raters that were slightly lower (0.02–0.10 SDs lower) than the ratings they received from African American raters. For comparison purposes, Sackett and Dubois noted that ratings for Caucasian ratees by African American raters ranged from 0.003 SDs lower to 0.003 SDs higher than ratings from Caucasian raters.

They concluded that race of the rater accounted for an extremely small amount of variance in the rating data.

Like ethnicity, the individual studies of age produced mixed results. The best available basis for drawing conclusions lies in available meta-analyses designed to look at the combined results from multiple studies. Two meta-analyses were found that investigated the effects of ratee age on performance ratings. McEvoy and Cascio looked at the combined results from 96 independent samples of the age–performance relation from 65 studies published between 1965 and 1986. They found that the mean correlation between age and objective measures of productivity (e.g., units produced, patents received, articles published) was 0.07. The relation between age and supervisor’s rating was similar (mean correlation = 0.03). Waldman and Avolio looked at the combined results from 37 samples reported in 13 research reports. Direct measures of productivity indicated that older workers were more productive. Older workers received lower ratings from their supervisors (average correlation = –0.14), but this relation was limited to nonprofessional workers. Based on the pooled results reported in these meta-analyses, there seems to be little relation between age and supervisors’ ratings for professional workers.

The relation between gender and performance ratings has received surprisingly little attention. The study by Pulakos et al. described earlier also investigated gender. Due to the very large sample, statistically significant differences were obtained, but Pulakos et al. noted the proportion of variance in ratings accounted for by gender was minimal. Male raters tended to rate women slightly more positively than men. Female raters on the other hand tended to rate males slightly more positively.

Only one study was identified that studied attractiveness and its relation to performance ratings.
Heilman and Stopeck\textsuperscript{91} conducted a laboratory study in which the attractiveness of the employee was varied experimentally by changing the picture on the employee’s written performance report. Participants in the study were MBA students who were asked to evaluate the performance and estimate advancement potential. Heilman and Stopeck\textsuperscript{91} concluded that attractiveness of the ratee inflated ratings of nonmanagerial women, deflated ratings of managerial women, and had no impact on the ratings of men. Because this was a single laboratory study with a very small number of participants rating hypothetical employees from written performance reports, the results should at most be considered suggestive of an area for further investigation.

### Impression Management

Certain practitioner characteristics are under personal control and can be manipulated to influence one’s ratings. Only two studies were found that investigated the effect of practitioner characteristics other than clinical competence on clinical performance ratings. Wigton\textsuperscript{92} studied case presentation style and its effect on the ratings assigned to medical students. He coached five first-year medical students to present each of five prepared (standardized) case presentations that varied systematically in content and organization. Fifteen experienced faculty raters evaluated videotaped records of the presentations. Results showed that the ranking given each presentation depended as much on which student was giving the presentation as the content of the presentation. Wigton concluded that the evaluation of clinical competence based on student case presentations could be significantly influenced by the personal characteristics of the students. This is the only experimental study of this type in medical education that we found. More experimental studies of this type would help us understand factors that influence clinical performance ratings. Kalet and colleagues\textsuperscript{24} conducted a study comparing faculty ratings on an OSCE to expert ratings. They concluded that faculty rated students on the basis of likeability rather than specific clinical skills.

There is a very active research community in social psychology investigating the processes by which people control the impressions others form of them dating back at least to the theoretical work of Goffman.\textsuperscript{93} Haas and Shaffir\textsuperscript{94,95} conducted investigations of some of these variables in their study of the professional socialization of one class of medical students over the period of their entire medical school training. Haas and Shaffir\textsuperscript{94,95} primary research methods included participant observation and interview research. They observed the students over the full range of their educational experiences. Referring to the sociological research on the process by which neophytes are socialized into professions, Haas and Shaffir\textsuperscript{94} noted that one characteristic is that neophytes must develop a “pretense of competence even though one may be privately uncertain” (p. 142). They proceeded to document the students’ development of this cloak of competence. Haas and Shaffir\textsuperscript{94} concluded by stating the following:

> A significant part of professionalization is an increased ability to perceive and adapt behavior to legitimators’ (faculty, staff, and peer) expectations, no matter how variable or ambiguous … In this context of ambiguity, students in both settings accommodate themselves, individually and collectively, to convincing others of their developing competence by selective learning and by striving to control the impressions others receive of them.” (p.148)

There have been many studies conducted in business organizations to determine the effects of subordinate’s use of impression management techniques on their supervisors’ ratings. There have also been a number of studies in the laboratory designed to control and separate the variables involved. All of these studies address concerns that supervisors’ evaluations of employees can be influenced substantially by factors other than job performance itself. The validity of performance ratings is compromised to the extent that this occurs. The results of these impression management studies are mixed, partly due to the variety of impression management techniques that have been studied and partly to the environmental conditions under which they have been studied. As a result, the best indication of the impact of impression management on performance ratings is found in the cumulative findings reflected in the meta-analysis conducted by Gordon.\textsuperscript{96} Gordon’s\textsuperscript{96} meta-analysis covered 55 journal articles, including 69 studies, yielding 106 effect sizes. Gordon\textsuperscript{96} found that impression management techniques studied so far appear to have little effect on performance ratings in field settings. There was a slightly larger effect in laboratory experiments, but the average effect was still characterized as small based on conventional standards for interpreting effect sizes. In summary, there is little hard evidence that impression management tactics are an important influence on supervisers’ ratings of performance.

### Influence by Other Evaluators

There are two traditional approaches to evaluating medical students and residents. In one, all supervisors are asked to fill out the rating form independently and submit it to the clerkship director or residency program director. The clerkship director or residency program director then reviews the individual ratings and assigns a final grade. The second approach requires that raters complete individual rating forms in preparation for a
group discussion of all practitioners by all supervisors. Each practitioner’s performance is discussed in committee and then a final rating is assigned through some consensus building process.

The two approaches have benefits and disadvantages. The consensus building model has the potential benefit of providing each rater with multiple perspectives on the practitioner’s performance and making all raters aware of data regarding practitioner performance in a variety of settings. However, many people express fear that such an approach may lead to a dominant individual unduly influencing the decision process. Although little data exist that address the benefits and weaknesses of the group model, the results of one study in the business literature is informative.

Martell and Borg97 conducted a laboratory study in which single individuals and groups of four rated the same performance either immediately or after a delay of 5 days. In the delayed rating condition, groups remembered performance details more accurately than individuals. However, groups also were more liberal in their rating criteria. On the other hand, Speer and colleagues78 found that grades assigned to medical students on an internal medicine clerkship were lower and closer to the faculty-stated appropriate average grade for clerks when the grades were assigned after a group discussion by faculty. These investigators attributed this difference to committee impartiality compared to individual faculty observers who worked with the students regularly on their internal medicine clerkship. Likewise, Hemmer et al.79 found that group performance assessment identified many professional behavior problems that were not otherwise disclosed. We found no studies to confirm or refute anecdotal reports that individuals with strong opinions or a persuasive style can sway group opinion inordinately in evaluative discussions. This is widely perceived as a disadvantage to the group process model and should be studied systematically.

Environmental Bias

As mentioned earlier, raters and practitioners interact in an environment with many distractions and confounding factors. Some of these factors are likely to influence judgments and ratings and are unrelated to true clinical competence. In this section, we discuss some of the confounding factors and provide evidence, where available, of their influence on performance ratings.

Time Pressures and Distractions

Anecdotal evidence and nonsystematic observation suggest that time pressures, distractions, and the pressure of competing responsibilities may compromise the ability of clinical supervisors to accurately observe and evaluate the clinical performance of physicians-in-training and practitioners. Time pressures and distractions likely influence the clinical performance assessment process most by decreasing the amount of observation that occurs and thus by decreasing the sampling of clinical performance across the domain of clinical situations and competencies of interest. They may also influence the rater’s time for reflection and the quality of information integration that underlies the assigned ratings. We found no research that addresses this question directly in either the clinical or general performance assessment research literature. One study in the decision-making literature is suggestive. Wright98 found that consumers, facing time pressures and distractions when making purchasing decisions, reduced search time by collecting fewer pieces of information and generally searching for negative information. It is possible that a similar process occurs in performance assessment. Clinical performance assessment practice would profit from more research in this area.

Clinical Performance = Group Effort

Patient care is a group effort in hospital settings and in most office settings. Many individuals including specialists, attendings, residents, and students participate in the process of working up, diagnosing, and managing patients. Therefore, it is difficult to assess the clinical competence of any one practitioner based on patient presentations and write-ups or on individual conversations about patients. Weaknesses in practitioner knowledge and skill may be masked by the contributions of others to the diagnosis and care of the patient as reported. One of the primary advantages of standardized patient encounters for clinical performance evaluation is the opportunity they give to assess an individual’s clinical proficiency knowing that the results reflect that individual’s clinical competence alone. We found no research that investigated the impact of this source of bias on traditional clinical performance ratings, nor did we find any reports of attempts to control this source of bias in field settings.

A great deal of research and evaluation on appraising team effectiveness99 in addition to individual effectiveness has occurred in the military, in business, and in the nuclear power industry in recent years. The medical profession has much to learn from this research legacy.

Effects of Being Observed

One factor that makes clinical performance assessment artificial is the fact that practitioners usually know when they are being observed. It is customary to think that we are seeing a person’s best performance under conditions where the person knows that he or she is being watched. However, it is also possible that the awareness of being observed creates anxiety and leads
to poor performance. In either case, appraisals may not provide an accurate reflection of how the individual will perform in practice when not under observation.

We found one study using clinical performance ratings by Kopelow et al.\(^{100}\) that addressed the effects of being observed. Kopelow and colleagues\(^{100}\) had practitioners work up the same standardized patient cases (portrayed by different individuals) both in a setting where they knew they were being evaluated and also in practice without evaluation awareness. One group of the practitioners saw the patients in the office first and then saw the patients under examination conditions. A second group took the examination first and saw the patients in their office later. The practitioners agreed to have the standardized patients rotate through their practice but were not made aware when the patients were introduced. Physicians recognized the standardized patients in only 26 of the 263 encounters (10%). Most pertinent to this article, average overall performance scores were 65% to 66% of possible points in the examination setting when practitioners knew they were being assessed and 52% to 56% in the practice setting when practitioners were not aware they were being assessed (\(p < .01\)).

Sackett et al.\(^{101}\) performed a comparable study in supermarkets. They instructed 1,370 supermarket cashiers from 12 supermarket chains to process (check out) 25 standardized items in a grocery cart while being observed. Cashiers were told that they would be timed and errors would be recorded. They were asked to do their best and to place equal emphasis on speed and accuracy. Each cashier processed six different standard cartloads of groceries over a period of 4 weeks. Sackett et al.\(^{101}\) considered these performance tests measures of maximum performance. All the cashiers worked in supermarkets where the cash register system kept records of items checked per minute, a measure of speed; and voids per shift, a measure of accuracy. These records provided a measure of typical performance. Sackett and colleagues\(^{101}\) established that the correlations between maximum and typical performance (speed and accuracy) were very low, leading them to conclude that the two measures do not yield comparable information about the relative performance of cashiers. They also found that supervisors’ ratings of the cashiers correlated more highly with the maximum than the typical performance measures confirming the sampling weaknesses noted for evaluation based on nonsystematic observation of critical incidents (positive or negative).

The results of the Kopelow et al.\(^{100}\) and Sackett et al.\(^{101}\) studies document the likelihood that performance under formal test conditions is not a very good indicator of “normal” performance. Furthermore, these results suggest that differences under these two sets of conditions are likely to be large. Their findings demonstrate the importance of creating ways to sample clinical performance in unobtrusive as well as more formal ways. Patient satisfaction measures\(^{102}\) and nurse observations\(^{103}\) may prove to be good ways of capturing information about normal clinical performance.

**Recommendations**

The following recommendations for improvement in clinical performance assessment practice are based on the research covered in this article.

**Broad, Systematic Sampling**

Practitioners need to be observed and evaluated across a broad range of clinical situations and procedures to draw reasonable conclusions about overall clinical competence. Further, it is unreasonable to expect that practitioners are equally effective across a broad range of situations. Efforts need to be directed toward assuring that practitioners can handle important, specific situations competently rather than a sole focus on identifying good and bad practitioners. Advanced Trauma Life Support (ATLS) and Advanced Cardiac Life Support training are good examples of training and assessment protocols designed to assure competence.

Medical educators need to be much more selective and deliberate in planning performance observations. Observations should cross a representative sample of chief complaints, diagnoses, tasks, and patient types. This will require planning and record keeping to fulfill the desired systematic sampling. Standardized patient encounters\(^9\)–\(^11\) and high-fidelity simulations\(^44\)–\(^46\) can be used to compensate for deficits in the naturally occurring patient mix.\(^104\) Educators also need to increase the absolute amount of observation that is done in support of clinical performance evaluation. One way to accomplish this may be to limit the amount of observation time during any one encounter. Results reported by Shatzer and his colleagues\(^{105,106}\) suggest that short periods of observation may be enough for clinical evaluation. A strategy that distributes available observation time across a larger number and range of encounters is preferable.

**Observation by Multiple Raters**

Although the primary goal should be to increase the range of situations and procedures observed, it is also the case that individual raters have idiosyncrasies regarding clinical focus (competencies they observe) and standards (stringency and leniency). Each practitioner should be rated by a number of different raters to balance the effects of these rater idiosyncrasies. The research that we have reviewed suggests that a minimum of 7 to 11 ratings should be collected to allow for a reproducible holistic estimate of general clinical competence. Ac-
quiring 7 to 11 ratings per practitioner with multiple raters, although not a small logistical task, appears to be feasible. Ramsey and colleagues sought to acquire 10 or more ratings for each of 228 practicing physicians. They reported that this was achieved for 72% of the physicians with no follow-up mailings.

**Keep Rating Instruments Short**

Our review indicates that raters view clinical competence as one or two dimensional. The work by Ramsey and colleagues suggests that a rating scale composed of 10 specific items plus a global item is sufficient when the goal is holistic decision making (i.e., about progress or promotion). Adding more items only boosts rater time and makes the task harder. Results of the Kreiter et al. study suggest that relatively little is gained in generalizability by adding more than 5 items to the rating scale when progress decision making is the goal. Careful instrument construction using a process similar to that used by Ramsey et al. should simplify the rating task with no loss of information. When the goal of evaluation is to provide feedback to support learning, the observation should be specific to the performance observed, and the feedback should be immediate and interactive. A general performance rating form is not appropriate for this purpose.

**Separate Appraisal for Teaching, Learning, and Feedback From Appraisal for Promotion**

One of the clearest implications from our review is that the feedback and progress decision goals of clinical performance evaluation should be separate. Physicians recall information regarding two clinical performance factors: (a) clinical skills, and (b) professional behavior. They do not recall performance details at the end of rotations when asked to fill out rating forms. Consequently, physicians should be encouraged to comment immediately about specific performance attributes that should be reinforced or corrected. This should include an interactive conversation about the situation, the practitioner’s behavior, desirable or undesirable consequences of the behavior, and some suggested alternatives if the behavior should be changed in the future. The implied message in performance assessment forms with room for written comments is that performance feedback should be saved and conveyed only through the form. We believe this is the wrong message to send.

**Encourage Prompt Recording**

Clinical performance evaluation usually involves completing an evaluation instrument at the end of a rotation based on observations that occurred during the rotation. As a result, there is some decrease in rating accuracy due to memory loss and distortion. Any method that encourages recording of observations and judgments at the time of the observation will minimize this problem. Encounter records can be referred to when completing end-of-rotation rating forms. The encounter records also provide a basis for providing detailed feedback to practitioners-in-training to help them improve their clinical performance. A number of approaches have been used to capture observations and judgments as they are made, including performance report cards (Rhoton and associates, Soglin, Brennan & Norman, diaries, and clinical evaluations filled out immediately).

**Supplement Formal Observation With Unobtrusive Observation**

Unobtrusive observation is valuable because it provides a better estimate of normal clinical performance than formal observation. To the extent that it increases the total number of clinical performance observations, it also helps to achieve the goal of broad sampling. Newble’s reported use of nurse evaluations and Tamblyn and associates use of patient evaluations are two effective ways of increasing unobtrusive evaluation.

**Consider Making Promotion and Grading Decisions Via a Group Review**

We believe that a group review process for making promotion and grading decisions has more benefits than problems associated with it. The group process provides each decision maker with a broader base of information on which to base the decision. It also provides added perspective because different raters tend to focus on different aspects of performance. Finally, available evidence suggests that groups are more likely to make unpopular promotion decisions than are individual raters making decisions in isolation. However, some physicians are reluctant to use a group process for fear that a person with a strong view will unduly sway the opinions of others. There is no evidence on this issue one way or the other.

**Supplement Traditional Clinical Performance Ratings With Standardized Clinical Encounters and Skills Training and Assessment Protocols**

Standardized clinical encounters have the advantage of allowing direct comparisons of practitioners when they perform exactly the same clinical tasks.
They also allow for evaluation of individual physicians’ skills in comparison to an established gold standard. Performance in these situations can also isolate the individual’s clinical performance capabilities and deficits. There is no contamination of the clinical performance measure resulting from the contributions of other members of the health care team who have cooperated in working up the patient. Standardized clinical encounters usually provide a more in-depth look at clinical performance. Finally, they provide a means of looking at the performance of an entire class (e.g., all third-year residents) and assuring that critical skills are acquired and maintained.

Educate Raters

At a minimum, make sure that all raters are familiar with the rating form at the beginning of the rotation. This will help raters focus their observations and organize performance information in a way that will facilitate performance assessment. Consider employing frame of reference training to assure that raters are calibrated regarding both definition of performance dimensions and performance quality rating category dimensions. More research is needed to determine whether adding other types of more extensive training will yield higher quality clinical performance ratings by physicians. The research by Newble and associates\(^8^5\) suggests that eliminating raters who are outliers (positive and negative) is a more cost-effective means of increasing the validity and reproducibility of clinical performance ratings than is training. It would be nice to have other studies investigate this practice to increase confidence in the reproducibility of the finding. We mention it here as much to encourage research as to suggest practice. We realize that the practice of culling raters may have undesirable consequences on faculty morale that need to be considered by the program or clerkship director, but the practice does merit consideration of the positive and negative consequences.

Provide Time for Rating

Do whatever you can to encourage intermittent observation and recording of clinical performance. Ratings and other types of clinical evaluation should be thoughtful, candid, and not rushed. Clinical evaluation of learners is an important professional activity that warrants protected time. We have heard anecdotally of departments that gather faculty members at a given time and place for the sole purpose of filling out performance ratings. The group rating process described earlier has the added benefit of protecting time for performance assessment.

Encourage Raters to Observe and Rate Specific Performances

The ABIM has explored use of the mini CEX. This process involves having an evaluator observe and rate specific performance on a single case immediately using an instrument tailored for that purpose. Given the problems of memory distortion associated with the normal rating process, we believe a supplemental process such as that proposed by the ABIM will provide additional, more accurate data about clinical performance that can be added to the mix when final promotion decisions are made. An even more optimal process would use rating scales specifically designed for selected chief complaints or procedures.

Use No More Than Seven Quality Rating Categories (i.e., 1 [Poor] to 7 [Excellent])

Five to seven quality rating categories are optimal. Even at this, the extreme rating categories will rarely be used. We also discourage two-level rating systems (e.g., 1–3 [unsatisfactory], 4–6 [satisfactory], to 7–9 [superior]). Additional quality rating categories add no performance information and just complicate the task for raters.

Establish the Meaning of Ratings

Given the tendency for raters to use primarily the top two categories and the reluctance to assign failing ratings, clinical performance assessment ratings only have meaning relative to each other. As with any measurement instrument, scores only acquire meaning through experience and systematic data analysis and interpretation. We encourage maintaining the same rating instrument and procedure. Clinical evaluators have more to gain by accumulating norms over a number of years than will be gained by “refining” the rating instrument. Familiarity with a form due to extensive use enables directors to easily and quickly identify outliers (i.e., those who perform unusually well or poorly based on their ratings). One can also develop a sense for the significance of particular ratings by keeping and using historical records regarding the ultimate status (e.g., asked to leave program, chose to leave program, ultimately judged to be outstanding) of practitioners with particular scores. This will require some investment to collect and make sense of the data, but it is the only way scores acquire meaning. Keeping the same form also makes it more feasible to conduct validity studies such as those conducted by Hojat and colleagues.\(^4^0\) Changing the form repeatedly does nothing but require recalibration of the new form, while losing data captured with an older, calibrated form.
Give Raters Feedback About Stringency and Leniency

Much like feedback to Olympic skating judges, feedback to raters about their clinical performance assessments may help calibrate future evaluations and increase the motivation to provide accurate ratings. Admittedly, because clinical raters observe different samples of clinical performance, they will have more difficulty interpreting information about stringency and leniency. However, providing feedback to raters regarding their average ratings along with average ratings for all raters provides some idea of rater stringency and leniency especially when the averages are based on many ratings. We are not convinced of the wisdom of handicapping raters arithmetically as rater knowledge that handicapping is being used may lead them to even more extreme ratings in the future.

Learn From Other Professions

Clinical medicine has much to learn about performance evaluation from other professions including the astronaut corps, aviation, business and industry, the military, and nuclear power plant operation. Each of these professions employs personnel evaluation theories and technologies that are applicable and useful in clinical medicine. They also have made more progress in evaluating team performance, an important clinical competency.

Acknowledge the Limits of Ratings

The consistent message throughout this target article is that clinical ratings, as an evaluation tool, have serious limits. Users of clinical ratings need to be fully aware of the limitations and make necessary adjustments in their use for making inferences about clinical competence and in making progress decisions. Despite recent arguments to the contrary, our review shows that faculty observations and the ratings they produce are insufficient measures of clinical competence by themselves although they provide an important source of information about clinical performance. The challenge ahead is to understand when and how to use traditional clinical ratings and to complement their use by adding other measures such as standardized clinical encounters and skills examinations (e.g., ATLS) to fill in the gaps regarding clinical performance.

References

Clinical Performance Ratings


Received 23 September 2002

Final revision received 14 April 2003