Increasing the Accuracy of Observer Ratings by Enhancing Cognitive Processing Skills

Diane E. Beck
School of Pharmacy, Auburn University AL 36849-5502

Patricia S. O'Sullivan
College of Nursing, University of Arkansas for Medical Sciences, 4301 West Markham Street, Little Rock AR 72205-7199

Larry E. Boh
School of Pharmacy University of Wisconsin, 425 North Charter Street, Madison WI53706

Ratings based on direct observation best extrapolate how the pharmacy student will perform in the actual practice setting and therefore, are a critical element in the overall assessment process. However, a review of the literature reveals observation-based ratings often suffer from poor rater accuracy and do not always measure how the student will perform in a variety of clinical situations. Teaching raters how to avoid the common rating errors and properly use a rating form have not resolved rater inaccuracy. This paper recommends giving attention to raters' cognitive processing skills and emphasizing the importance of frequent observations. The Cognitive Processing Model described in this paper can make experiential instructors more aware of how they acquire, store, recall, and integrate information into ratings. Pharmacy schools are encouraged to provide their experiential faculty with a rater training program that emphasizes this model.

INTRODUCTION

In 1976, Elenbaas(1) described well-conceived mechanisms for evaluating PharmD students using examinations and patient cases. His paper included a rating form developed to evaluate a student's ability to manage a patient case in a conference setting. He also described use of a quarterly written exam where students had to achieve a minimum score established from the mean score achieved by program faculty. A second multiple-choice examination provided a comparison of the student's performance and that of their peers. In addition, an oral examination was used as another means of evaluating therapeutic knowledge.

Although program faculty felt these assessment methods were effective, Elenbaas questioned whether they predicted performance in the actual clinical setting. Elenbaas recognized the difficulty in accomplishing this, but stressed the importance of further work in this area. He also cited the need to assess whether these assessment methods were accurate.

Since publication of this paper almost 20 years ago, numerous pharmacy educators have described use of examinations or written simulations to assess experiential student performance(2-14) However, only three have reported the reliability of the described methods1(2-14). No one has published data confirming that their assessment methods predict performance in the actual clinical setting. Although educational researchers believe ratings based on direct observations best predict performance in the routine practice setting(15), pharmacy educators have reported neither the reliability nor the validity of observation-based ratings. This is particularly distressing since a recent survey of experiential pharmacy programs revealed our students are frequently evaluated by observational methods(16).

With the movement to entry-level PharmD degree programs, pharmacy educators have seen an increase in credit hours allocated to experiential course-work and the need for quality assessment methods. In this paper, principles supporting the importance of observation-based performance ratings in experiential education will be established. Rater accuracy is the greatest limitation of observation-based ratings and this problem has been attributed to raters' cognitive processing skills. Therefore, the Cognitive Processing Model will be introduced as a new framework for rater training programs. This model promotes greater awareness of how a rater acquires, stores, recalls, and integrates information into ratings.

Because performance appraisal terminology may be new to some pharmacy educators, key definitions are provided in the glossary. Several definitions require special emphasis at this point. As defined in the glossary, the terms reliability and accuracy infer different connotations. Interrater reliability is evident when two raters observe a student and independently assign similar ratings. However, if both of these raters bias their ratings in the same direction, they may agree on the assigned ratings, but their scores are not indicative of the "true performance." Accuracy therefore, infers the ratings exhibit interrater agreement and in addition, are a measure of the true performance. Before reading further, the reader should also note the definitions of assessment and evaluation (see Appendix). In the education literature, assessment and evaluation are frequently used interchangeably. However in this paper, the term "evaluation" will be reserved for discussions related to inferences of validity.

Table I. Inferences of validity for evaluation methods frequently used to assess professional competence

	Inferences on validity		
Evaluation method	Evaluation[a]	Generalization[b]	Extrapolation[c]
Simulations (OSCE)	+	+	?
Written objective examinations	+	+	0
Oral examinations	±	±	0
Direct observations	±	±	+
Patient presentations	+	0	0

0 = Weak; ±=Sometimes weak; +=Strong
[a]The assessment method correctly differentiates good from poor performance.
[b]The assessment method can correctly predict performance in a variety of situations based on only a sample of test items.
[c]The assessment method predicts the student's performance in routine practice.
Adapted from Kane(15).

IMPORTANCE OF DIRECT OBSERVATIONS

To substantiate the importance of observation-based ratings, we will first examine the concept of test validity. The concept of test validity has been evolving over the last 10 years(17,18). As in the past, an assessment method is considered valid if it measures what it was intended to measure(19,20). One establishes evidence of validity by accumulating inferences supporting this premise. In the past, educators established validity by reporting either content, criterion, or construct validity. However, measurement researchers now assert there are a variety of ways to accumulate evidence inferring validity and an assessment method should exhibit several types. The most appropriate types of evidence should be based on the professional judgment of those who administer and interpret the test/assessment results(21-23).

For example, Kane(15) has proposed that three types of inferences are important in establishing the validity of assessment methods in medical education. The inferences cited by Kane are: (*i*) evaluation; (*ii*) generalization; and (*iii*) extrapolation. An inference of evaluation indicates the assessment method can accurately differentiate a good from a poor performance. The inference of generalization refers to the correct prediction of performance in a variety of practice situations based on only a sample of test items or observations during the assessment period. Extrapolation, the third inference posited by Kane, indicates that the test score or rating predicts how the student will perform in the actual practice setting.

Using the inferences of evaluation, generalization, and extrapolation as a framework. Kane examined the strengths and weaknesses of assessment methods frequently used to establish professional competence(15). Kane critiqued: (*i*) simulations; (*ii*) objective tests; and (*iii*) direct observation of performance. Of these three assessment methods, direct observation of daily performance best predicts how an individual will perform in practice. Table I summarizes Kane's assessment. As summarized below, he gave special attention to each method's weakest inferences of validity since all of these methods have weaknesses.

Kane notes that during simulation testing (*e.g.*, Objective Structured Clinical Examinations or OSCEs), instructors can accurately score and predict performance in a variety of practice situations(24-33). As a result, the inferences of evaluation and generalization are usually strong. However, educators lack evidence that OSCEs predict student performance in the actual patient care setting(15,33,34).

With respect to written exams, Kane acknowledges instructors usually achieve accurate and reliable scoring; therefore, the inference of evaluation is strong(15). Objective tests have the capacity to sample a wide range of content making the inference of generalizability strong. However, extrapolation is a weak inference because such testing is, at best, an indirect measure of what happens in practice. For example, written examinations often fail to assess problem-solving and/or communication skills with patients. Although the inference of extrapolation is greater with oral examinations, they still do not predict performance in routine practice. Accurate scoring is also more difficult to achieve with oral examinations.

Because direct observations of daily performance take place in the actual practice setting. Kane posits their inference of extrapolation is strong. However, the inferences of generalization and evaluation are weak. Generalization suffers because competing demands such as patient care, meetings, teaching, and research limit the instructor's opportunities to make observations. Further, at some experiential sites the student may not encounter a variety of tasks. Therefore, even if the instructor has time to observe the student, the observations will not be of sufficient variety to establish generalizability.

Ratings based on direct observations of daily performance frequently exhibit inaccuracy and the etiology of this is multifactorial(35-38). For example, delays between observing and rating a student's performance can result in distortion and bias(39,40). The rater's cognitive capacity limits the amount of information one can process when assessing performance involving complex behaviors and decisions(41). Also, the instructor must consider instances when the student's management of a patient may depend on the physician's decisions or circumstances beyond the student's control. The pervasiveness of these factors and degree of raters' awareness of them often weakens the link between observation-based performance ratings and the inference of evaluation.

Another observation-based assessment method used to infer student performance in the practice setting is the formal patient or case presentation. During such a presentation, the student presents a patient case to a group of peers and faculty and describes relevant literature(42). However, patient presentations have weak inferences of extrapolation and generalization. Since the instructor is not observing the actual patient event, it is difficult to extrapolate how the student performs in routine practice. The inference of generalization is weak because the case represents management of a single patient. It is difficult from this single case to generalize the student's ability to manage multiple patients who have problems of varying complexity. Like other observation-based assessment methods, the accuracy of patient presentation ratings is often poor due to factors such as the rater's limited cognitive capacity, and distortion and bias of observations.

In summary, trade-offs occur when inferring validity of either objective exams, direct observation of performance, patient presentations, or simulations (see Table I). Faculty can achieve greater confidence in decisions about pharmacy student performance by using a combination of assessment methods. Since the inference of extrapolation is strongest with observation-based ratings, this assessment method should be the major component of a school's clinical evaluation system. To make observation-based ratings valid, the two weaknesses we have delineated must be remedied. First, to overcome the problem of an insufficient number and type of observations, faculty must make time for observing the student. Second, research by behavioral scientists provides instructors useful strategies for enhancing rater accuracy.

PERFORMANCE ASSESSMENT RESEARCH

Over the last 60 years, most research concerning performance assessment has focused on enhancing interrater reliability by either improving the instrument or the rater(43-45). Studies have attempted to improve rating reliability by designing better rating instruments. However, studies have shown minimal differences in random error and reliability among the various formats such as Behaviorally Anchored Rating Scales and Behavioral Observation Scales(46-48). While good assessment forms are essential, there is a limit to what forms alone can do.

Some investigators have focused on the rater to improve the assessment of performance. For example, researchers have evaluated Rater Error Training (RET), which concentrates on avoiding the common rating errors such as halo, leniency/stringency, and central tendency(49,50). The literature frequently cites avoidance of these errors enhances performance rating reliability. However, rater training addressing only these errors has not increased the accuracy of ratings (*i.e.*, a true measurement of the performance)(51). Rater Accuracy Training (RAT), which emphasizes appropriate use of the rating instrument, improves rater accuracy but, has not completely rectified the problem of rating errors(43,50).

In 1980, Landy and Farr(43) reviewed studies that had focused on traditional rater training and rating scales and concluded further research in these two areas would not resolve all causes of systematic and random rating error. Other researchers have subsequently agreed that to increase rating accuracy, we must gain greater insight into how raters acquire, encode, retrieve, and integrate the information into a rating(44,52,53). These four cognitive processing skills represent the human "black-box" of the performance assessment process(41).

A number of researchers have proposed cognitive models of rating performance(41,44,53-56). These various models differ only in the emphasis they place on the four cognitive processes described in Figure 1(44).

Data supporting cognitive processing models stem primarily from laboratory studies(57,58). One advantage of laboratory-based studies is that the researcher can control the subject's performance and therefore, measure the true performance. This allows for determination of rating accuracy and not just reliability. Unfortunately, results from laboratory experiments may not represent the "real world." However, there is some evidence, based on a recent study conducted under both laboratory and realistic conditions, that laboratory-based performance studies correlate with those conducted in a more realistic setting(58).

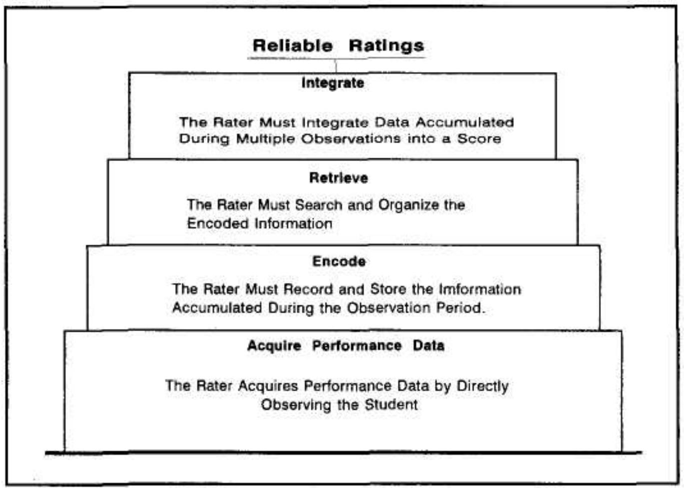
Reliable Ratings
Integrate
The Rater Must Integrate Data Accumulated During Multiple Observations into a Score
Retrieve
The Rater Must Search and Organize the Encoded Information
Encode
The Rater Must Record and Store the Imformation Accumulated During the Observation Period.
Acquire Performance Data
The Rater Acquires Performance Data by Directly Observing the Student

Fig. 1. The Cognitive Processing Model. Mental processing of performance data requires the rater to acquire, encode, retrieve and integrate performance information.

THE COGNITIVE PROCESSING MODEL

The Cognitive Processing Model depicted in Figure 1 delineates the four cognitive processes experiential instructors should strengthen to improve their assessment of pharmacy students. According to the Cognitive Processing Model, the rating task is a four-stage process that requires the rater to acquire, encode, retrieve, and integrate performance data into scores. Since each plays a critical role in the rating process, we will discuss them in detail.

Acquisition. The rater must first acquire information via observation of the student. This acquisition requires the rater to invoke either automatic or controlled processing of what is observed(53,59,60). In addition, the information acquired by a rater depends on knowledge of the rating instrument, time pressures placed on the rater, and the rater's preconceived notions(54).

Researchers believe that acquisition of information occurs by automatic and controlled processes. Each process is routinely invoked by raters, but both have limitations. Therefore, the rater must be trained on how to minimize the limitations of each.

Automatic processing occurs when the student's performance meets the instructor's expectations(61). During automatic processing, the instructor recognizes and stores the observed behavior automatically (without conscious awareness). However, cues such as sex, dress, and speech may unknowingly influence the categorization of a student during automatic processing. This initial categorization may have consequences for how the instructor categorizes future behaviors and recalls observations when determining scores. The instructor must therefore, maintain vigilance that this bias may occur.

When the behaviors do not fit the categories used during automatic processing or the relevancy of the task is questionable, the rater engages controlled processing. Since controlled processing involves observation of unusual behaviors, the rater must assess first their significance. If significant, the rater must then establish categories for these behaviors. This processing requires the rater's attention and is limited by the rater's cognitive capacity(54).

Attribution theory best explains how the rater determines the most appropriate category during controlled processing(54). According to attribution theory, the rater attributes the behavior to either: (*i*) some force in the environment (*i.e.*, ineffective medical team); (*ii*) chance (*i.e.*, most students would not know that answer); or (*iii*) the student (*i.e.*, ability or effort). This decision represents one potential source of random error that can lead to rating inaccuracy. For example, Feldman has noted that raters probably attribute performance to trainee ability too often and underestimate forces in the environment(54).

If the rater attributes the behavior to the student, the observation must then be categorized according to the level of performance. McDonald(62) demonstrated that an orientation to the dimensions of the rating instrument prior to doing actual observations improves the accuracy of this categorization. Raters who received either no information or misinformation about the dimensions produced less accurate ratings. Poor accuracy likely occurred because they paid attention to information differently. McDonald proposed the instrument's dimensions served as the categories for behaviors to which the rater should attend. These dimensions helped raters differentiate relevant from irrelevant observations. The raters paid no attention to nonpertinent information and therefore, did not store or retrieve it.

Kamouri and Balzer(63) determined that when time pressures limit the number of observations, ratings will become less accurate. They found raters who made more observations of performance assigned more accurate ratings. Also, if several subjects, with differing levels of performance, were rated simultaneously, rating accuracy was highest when raters observed all of them equally.

Balzer(64) found a rater's observations can be influenced by an earlier interaction or a prior assessment by another individual. This investigator established that raters were more likely to record behavioral incidents that did not agree with initial impressions than those that did. In summary, there are identifiable influences on the acquisition of performance information. Raters need to attend to these influences since evidence suggests that accuracy of observations correlates with accuracy of performance ratings(65).

Encoding (Record and Store). Raters can enhance their cognitive processing by effectively recording and storing the information gained from an observation. These two encoding skills can be enhanced by either improving memory or maintaining a diary. However, maintenance of a diary is most valuable because it provides a "hard copy" of the incidents supporting the ratings assigned(66).

DeNisi *et al.* posit some raters can organize information in memory(66). Raters who can encode observations using memory alone typically are more experienced. DeNisi and colleagues suggest that raters who successfully rely on memory may have developed their own strategies for organizing information or developed a greater memory capacity. Since these skills require practice, inexperienced raters are less likely to have this ability. Therefore, new instructors particularly need a strategy to encode their observations into memory(66).

Several researchers(66-68) have demonstrated that maintenance of a written record of observations can improve the rater's recall of observations. These findings support Guion's(69) earlier hypothesis that, to enhance rating accuracy, the rater should record both good and bad tasks completed by the subject being evaluated. The rater can then retrieve these recorded observations at the time of assessment. DeNisi *et al.* (66) found that those who use a diary are better able to recall incidents than those not using one. In addition, those who use diaries organized by subject recall and rate more incidents correctly.

When Guion proposed the concept of diary-keeping, he theorized that diaries improve rater recall at the time of rating performance(69). Others have suggested that diary keeping results in deeper encoding at the time of observation, better structuring of information in memory, or more critical assessment of behaviors during observation(68). However, the mechanism by which diary keeping improves rating accuracy is most likely multifaceted. Larson found that when rating performance, raters store in memory two different types of information about a subject(68). First, they store a semantic description of the subject's behaviors. Second, raters make stimulus-based judgments meaning they assign a rating to each specific incident during the observations; at the time of assigning ratings, these individual ratings are mentally averaged. Larson has proposed stimulus-based judgments are of higher quality because the rater evaluates the behavior more critically while writing down the information.

In summary, researchers have shown that diary keeping enhances the encoding of observations. Raters who believe they can successfully encode their observations using memory alone, should assess their ability to do this before assuming they have this skill. However, diary keeping is preferable since it provides the instructor with specific data should a student contest a grade.

Retrieval. Just prior to assigning a score, the instructor needs to accurately retrieve the encoded information. Accurate retrieval requires the instructor to search the encoded information and assimilate it in an organized manner.

Srull *et al.* (70) found that raters can recall more easily information that is stored initially in a clear pattern. Because of these research findings, Cafferty *et al.* (71) conducted two studies to determine how raters search, organize, and use performance information about multiple individuals. These researchers found that most raters examine a series of one person's performances before moving on to the next person being evaluated. In addition, searching and organizing by persons results in raters recalling more items than does searching and organizing by tasks. Cafferty and colleagues(71) also found raters can accurately rate each subject's overall performance (assign a single overall score) no matter what method they use. However, when raters retrieve and organize by tasks, they can more accurately differentiate performance among those being rated. Based on these results, Cafferty *et al.* (71) recommend that although raters recall less information, they can get more accurate ratings of several subjects by retrieving written incidents that are organized by task and then subject.

However, several factors may compromise recall. First, recall is complicated because raters encode some facts for one purpose and must reprocess them before they can evaluate them for another purpose(72). Stressful work conditions can also affect the retrieval and possibly input of information. Srinivas and Motowidlo(73) indicate that raters, under stressful conditions, depend more on their initial impressions of the subject and differentiate less among the performance dimensions(a halo effect). Raters must be cognizant of these factors so that they do not interfere with the assignment of a score.

Integration of Information. The integration of information into a score comprises the last and most complex step of cognitive processing. Two theories propose how a rater integrates information into a score. The first assumes that the rater evaluates each individual incident during the observation. Then during integration and completion of the instrument, the rater uses some weighting schema to derive a single rating for each performance dimension (stimulus-based judgments)(74,75). The second theory, known as "cognitive categorization," presumes that as the rater makes observations, the rater continuously integrates them into global categories of performance. Then, at the end of the observation period, the rater assigns a score that corresponds to this general impression(75).

The results of a study by Nathan and Lord are more supportive of stimulus-based judgment theory in which the rater evaluates each individual incident and, at the time of assigning a score, uses a weighting schema to derive a rating(75). However, because of a large halo effect, these investigators have concluded raters also integrate observations into global categories of performance (cognitive categorization).

At the end of the assessment period, the rater must integrate information from multiple observations into a single score for each dimension. This integration is particularly difficult because student performance usually fluctuates. Research findings from social psychology suggest that when the rater has to integrate findings from multiple observations into a score, they weigh the initial (primacy) and most recent (recency) observations more than those that occurred during the middle of the assessment period(76,77). Two possible hypotheses support the occurrence of primacy and recency effects. According to the attention decrement hypothesis, people place greater weighting on early information and less on the most recent facts. Alternatively, the consistency hypothesis proposes that first impressions change as new information accumulates.

Steiner *et al.* conducted a study to evaluate the hypotheses that support primacy and recency effects(78,79). In their study, 333 psychology students viewed four videotaped lectures. One lecture represented either a poor or good performance and the other three lectures showed an average performance. The researchers manipulated the order of the lecturers' inconsistent performance among the students who served as raters. The students first rated the performance of the lecturers in a single session. During the second part of the study, the subjects observed these four lectures over four days.

Steiner and colleagues found that when the inconsistent performance was poor and occurred in the most recent lecture, the overall ratings were biased as poor. The investigators termed this a "recency effect." Steiner and colleagues also concluded that inconsistent performance influences overall impressions. A finding that overall ratings were weighted in the direction of the inconsistent poor performance is supported by research on interviewer judgment. Specifically, Schmitt(80) found that, during employment interviews, interviewers weighted negative information more heavily than positive.

Several studies have shown that liking interferes with the processing of information(81,82). Liking is an affect where there is interpersonal attraction. This emotion can also go in the opposite direction when an individual "just dislikes" someone. Zalesny and Highhouse(83) have shown that when school administrators perceived their student teachers had teaching attitudes similar to theirs, they rated the student more accurately than when their perceptions contradicted each other. Liking/disliking interferes with the processing of information and they believe that training may reduce it.

Enhanced cognitive processing skills can strengthen an instructor's ability to rate with accuracy; however, the instructor must be willing to assign the most appropriate score(58). DeCotis and Petit cite literature indicating that raters are more willing to give accurate ratings when the scores are intended for counseling and feedback rather than administrative decisions such as merit or promotion(84). Experiential educators should consider the willingness to rate accurately when comparing formative and summative assessments of pharmacy student performance.

DISCUSSION

The most important consideration when assessing student performance in the experiential setting is extrapolating how the student will perform in routine practice. Of the assessment methods described in this paper, direct observations of daily performance best infer this. However, ratings assigned by raters are frequently inaccurate. In order to increase the accuracy of these ratings, raters must enhance their cognitive processing of performance data. The Cognitive Processing Model described in Figure 1 provides all health profession educators with a framework for enhancing the accuracy and therefore, reliability of observation-based performance ratings.

Most individual instructors do not have the time or knowledge base to extrapolate evaluation strategies from performance evaluation literature. Therefore, pharmacy schools are encouraged to develop or gain access to training programs for their faculty that emphasize the Cognitive Processing Model. Such rater-training programs will likely be more effective than those traditionally used in health professions education(35-38). Educators in medicine and nursing have described rater training programs when reporting the reliability of observation-based rating instruments. However, in most instances, the content of the rater training program was either not described or consisted of Rater Error Training (RET). RET has been most frequently described by medical educators. This type of training program emphasizes avoidance of rating errors such as halo and does not teach strategies to enhance rating accuracy. Reliance on only this technique may explain one reason why rating reliability and/or accuracy in medical education has been disappointing(35-38).

Although nursing educators have noted use of rater training, none have described the program used in enough detail to discern its qualities(85-87). These studies have reported acceptable reliability however, they were pilot studies that involved only a few selected nurses as raters. In our review of the medical literature, we found only one reference suggesting that better cognitive processing may enhance rater accuracy and reliability. Littlefield(88), who has extensive experience in the assessment of medical student and resident performance, briefly cited literature indicating the importance of cognitive processing. However, no one has since reported use of rater training programs that emphasize enhancement of cognitive processing skills.

The most important reason for having experiential faculty observe student performance is so they can supervise (*i.e.*, teach) the student. The instructor can most effectively help the student improve his/her patient care skills by making observations, and assimilating this information into feedback. Therefore, teaching in the practice setting also requires understanding and use of effective cognitive processing skills. Enhanced cognitive processing of observations should therefore, make our faculty better teachers.

SUMMARY

Of the commonly used methods to evaluate experiential students, ratings of daily performance based on observations continue to best extrapolate how the student will perform in routine practice. Instructors can achieve greater confidence in these ratings by making a sufficient number of observations and increasing rater accuracy. Rating scale formats and traditional forms of rater training have not resolved all causes of systematic and random rating error. To further reduce these errors, clinical faculty must become more aware of how they acquire, encode, retrieve, and integrate performance information into a rating. The Cognitive Processing Model provides pharmacy educators a framework for developing rater-training programs that can achieve this.

Am. J. Pharm. Educ., **59**, 228-235(1995); received 4/18/94, accepted 6/8/95.

References

(1) Elenbaas, R.M., "Evaluation of students in the clinical setting," *Am. J. Pharm. Educ.*, **40**, 410-417(1976).

(2) Smith, H.A. and Kifer, E., "Concepts, models and methodologies for the evaluation of experiential education in pharmacy." *ibid.*, **42**, 159—166(1978).

(3) Andrew, B.J., "A technique for assessment of pharmacy students skills in patient interviewing," *ibid.*, **37**, 290-299(1973).

(4) Warner, D., "Evaluation of a pilot pharmacy clerkship." *ibid.*, **34**, 256-264(1970).

(5) Speranza, K.A., "A diary study of the socialization process in a hospital clinical experience course," *ibid.*, **39**, 24-30(1975).

(6) Tobias, D.E., Speedie, S.M. and Kerr, R.A., "Evaluation of clinical competence through written simulations," *ibid.*, **42**, 320-323(1978).

(7) Nelson, A.A., Bober, K.F. and Bashook, P.G., "Evaluation of a college-structured practical experience program," *ibid.*, **40**, 232-236 (1976).

(8) McKenzie, M.W., Johnson, S.M. and Bender, K.J., "A competency-based, self-instructional module on medication history interviewing for pharmacy students: Rationale, description and formative evaluation," *ibid.*, **41**, 133-142(1977).

(9) Tobias, D.E., Michocki, R.J. and Edmondson, W.H., "Evaluation of students in a competency-based undergraduate clinical clerkship," *ibid.*, **42**, 31-34(1978).

(10) Carter, R.A., Bennett, R.W., Black, CD. and Blank, J.W., "Use of simulations as performance standards for external degree doctor of pharmacy students," *ibid.*, **49**, 119-123(1985).

(11) Friend, J.R., Wertz, J.X., Hicks, C. and Billups, N.F.,"A multi-faceted approach to externship evaluation," *ibid.*, **50**, 111-126(1986).

(12) Roffman, D.S., Tobias, D.E. and Speedie, S.M., "Validation of written simulations as measures of problem solving for pharmacy students," *ibid.*, **44**, 16-24(1980).

(13) Pancorbo, S., Holloway, R.L., McNeiley, E. and McCoy, H.G.," Development of an evaluative procedure for clinical clerkships," *ibid.*, **44**, 12-16(1980).

(14) Guidry, T.D. and Cohen, P.A., "A practical examination for student assessment in an externship program." *ibid.*, **51**, 280-284(1987).

(15) Kane, M.T., "The assessment of professional competence," Eval. *Health Prof.*, **15**, 163-182(1992).

(16) Boh, L.E., Pitterle, M.E., Schneider, F. and Collins, C.L., "Survey of experiential programs: course competencies, student performance and preceptor/site characteristics," *Am. J. Pharm. Educ.*, **55**, 105-113(1991).

(17) Messick, S., "Meaning and values in test validation: The science and ethics of assessment," *Educational Researcher*, **18**, 5-11(1989).

(18) Cronbach, L.J., "Five Perspectives on Validity Argument," in *Test Validity*, (edits. Wainer, H. and Braun, H.I.) Lawrence Erlbaum Associates, Publishers, Hillsdale NJ (1988) pp. 3-17.

(19) American Educational Research Association, American Psychological Association, and National Council on Measurement, *Standards for Educational and Psychological Testing*, American Psychological Association, Washington DC (1985) pp. 9-18.

(20) Angoff, W.H., "Validity: An Evolving Concept," in *Test Validity*, (edits. Wainer, H. and Braun, H.I.) Lawrence Erlbaum Associates, Publishers, Hillsdale NJ (1988) pp. 19-32.

(21) Messick, S., "The once and future issues of validity: Assessing the meaning and consequences of measurement," in *Test Validity*, (edits. Wainer, H. and Braun, H.I.) Lawrence Erlbaum Associates, Publishers, Hillsdale NJ (1988) pp. 33-44.

(22) Risucci, D.A. and Tortolani, A.J., "A methodological framework for the design of research on the evaluation of residents," *Acad. Med.*, **65**, 36-41(1990).

(23) Messick, S., "Validity," in *Educational Measurement*, (edit. Linn, R.L.) Third Edition. American Council on Education and MacMillan Publishing Company, New York NY (1989) pp. 13-103.

(24) van der Vleuten, C.P.M. and Swanson, D.B., "Assessment of clinical skills with standardized patients: State of the art," *Teach. Learn. Med.*, **2**, 58-76(1990).

(25) Stillman, P.L., Swanson, D.B., Smee, S., et al., "Assessing clinical skills of residents with standardized patients," Ann. Intern. Med., **105**, 762-771(1986).

(26) Vu, N.V., Barrows, H.S., Marcy, M.L., *et al.*, "Six years of comprehensive, clinical, performance-based assessment using standardized patients at the Southern Illinois University School of Medicine," *Acad. Med.*, **67**, 42-50(1992).

(27) Anderson, M.B., Stillman, P.L. and Wang, Y., "Growing use of standardized patients in teaching and evaluation in medical education." *Teach. Learn. Med.*, **6**, 15-22(1994).

(28) Tamblyn, R., Abrahamowicz, M., Schnarch, B., *et al.*, "Can standardized patients predict real-patient satisfaction with the doctor-patient relationship?," *ibid.*, **6**, 36-44(1994).

(29) Colliver, J.A., Marcy, M.L., Vu, N.V., *et al.*, "Effect of using multiple standardized patients to rate interpersonal and communication skills on intercase reliability," *ibid.*, **6**, 45-48(1994).

(30) Shatzer, J.H., Wardrop, J.L., Williams, R.G. and Hatch, T.F., "Generalizability of performance on different-station-length standardized patient cases," *ibid.*, **6**, 54-58(1994).

(31) Barrows, H.S., "An overview of the uses of standardized patients for teaching and evaluating clinical skills," *Acad. Med.*, **68**, 443-451(1993).

(32) Colliver, J.A. and Williams, R.G., "Technical issues: Test application." *ibid.*, **68**, 454-460(1993).

(33) Miller, G.M., "Commentary on assessment of clinical skills with standardized patients: State of the art," *Teach Learn. Med.*, **2**, 77-78(1990).

(34) Friedman, M. and Mennin, S.P., "Rethinking critical issues in performance assessment," *Acad. Med.*, **66**,.390-395(1991).

(35) Noel, G.L., Herbers, J.E., Caplow, M.P., *et al.*, "How well do internal medicine faculty members evaluate the clinical skills of residents?" *Ann. Intern. Med.*, **117**, 757-765(1992).

(36) O'Donohue, W.J. and Wergin, J.F., "Evaluation of medical students during a clinical clerkship in internal medicine," *J. Med. Educ.*, **53**, 55-48(1978).

(37) Littlefield, J.H., Harrington, J.T., Anthracite, N.E. and Garman, R.E.," A description and four-year analysis of a clinical clerkship evaluation system," *ibid.*, **56**, 334-340(1981).

(38) Thompson, W.G., Lipkin, M., Gilbert, D.A., *et al.*, "Evaluating evaluation: Assessment of the American Board of Internal Medicine resident evaluation form," *J. Gen. Intern. Med.*, **5**, 214-217(1990).

(39) Cooper, W.H., "Conceptual similarity as a source of illusory halo in job performance ratings," *J. Appl. Psychol.*, **66**, 302-307(1981).

(40) Murphy, K.R. and Balzer, W.K., "Systematic distortions in memory-based behavior ratings and performance evaluations," *ibid.*, **71**, 39-44(1986).

(41) Wexley, K.N. and Klimoski, R. "Performance appraisal: An update,"

in, *Performance Evaluation, Goal Setting*, *and Feedback*,(edits. Ferris, and Rowland, K.M.) JAI Press, Inc., Greenwich CT (1990) pp. 1-45.

(42) Beck, D.E. and Clayton, A.G., "Validity and reliability of an instrument that evaluates PharmD student performance durins a patient presentation," *Am.J. Pharm. Educ*., **54**, 268-274(1990).

(43) Landy, F.L. and Farr, J.L., "Performance rating," *Psychol. Bull*., **87**, 72-107(1980).

(44) DeNisi, A.S. and Williams, K.J., "Cognitive approaches to performance appraisal," in *Performance Evaluation*, *Goal Setting*, *and Feedback*, (edits. Ferris, G.R. and Rowland, K.M.) JAI Press, Inc., Greenwich CT (1990) pp. 47-93.

(45) IIgen, D.R. and Feldman, J.M., "Performance appraisal: A process focus," *Res. Organizational Behav*. **5**, 141-97(1983).

(46) Borman, W.C. "Behavior-based rating scales," in *Performance Assessment-Methods and Applications*, (edit. Berk, R.A.) Johns Hopkins University Press, Baltimore MD (1986) pp. 100-120.

(47) Borman, W.C. "Format and training effects on rating accuracy and rater errors," *J. Appl. Psychol*., **64**, 410-421(1979).

(48) Jacobs, R.R. "Numerical rating scales," in *Performance Assessment-Methods & Applications*, (edit. Berk, R.A.) Johns Hopkins University Press, Baltimore MD (1986) pp. 82-99.

(49) Latham, G.P., Wexley, K.N. and Pursell, E.D., "Trainingmanagers to minimize rating errors in the observation of behavior." *J. Appl. Psychol*., **60**, 550-555(1975).

(50) Pulakos, E.D., "A comparison of rater training programs: Error training and accuracy training," *ibid*., **69**, 581-588(1984).

(51) Borman, W.C, "Individual differences correlates of accuracy in evaluating others' performance effectiveness," *Appl. Psychol. Meas*., **3**, 103-15(1979).

(52) Landy, F.J. and Farr, J.L., *The Measurement of Work Performance-Methods*, *Theory*, *and Applications*, Academic Press, New York NY (1983) pp. 22-23.

(53) Wherry, R.J. and Bartlett, C.J., "The control of bias in rating," *Personnel Psychology*, **35**, 521-551(1982).

(54) Feldman, J.M., "Beyond attribution theory: Cognitive processes in performance appraisal," *J*., *Appl. Psychol*., **66**, 127-148(1981).

(55) DeNisi, A.S., Cafferty, T.P. and Meglino, B.M., "A cognitive view of the performance appraisal process: A model and research propositions," *Organizational Behavior and human Performance*, **33**, 360-96(1984).

(56) Cooper, W.H., "Ubiquitous halo," *Psychol. Bull*., **90,** 218-244(1981).

(57) Banks, C.G. and Murphy, K.R., "Toward narrowing the research-practice gap in performance appraisal," *Personnel Psychol*., **38**, 335-345(1985).

(58) Roach, D.W. and Gupta, N., "A realistic simulation for assessing the relationships among components of rating accuracy," *J. Appl. Psychol*., **77**, 196-200(1992).

(59) Cronbach, L.T., "Processes affecting scores on 'understanding of others' and 'assumed similarity,'" *Psychol. Bull*., **52**, 177-93(1955).

(60) Krzystofiak, F., Cardy, R. and Newman, J., "Implicit personality and performance appraisal: The influence of trait inferences on evaluations of behavior." *J. Appl. Psychol*., **73**, 515-21(1988).

(61) Schneider, W. and Shiffrin, R.M., "Controlled and automatic human information processing: I. Detection, search, and attention," *Psvchol. Rev*., **84**, 1-66(1977).

(62) McDonald, T., "The effect of dimension content on observation and ratings of job performance," *Organizational Behavior and Human Decision Processes*, **48**, 252-71(1991).

(63) Kamouri, A.L. and Balzer, W.K., "The effects of performance sampling methods on frequency estimation, probability estimation and evaluation of performance information," *ibid*., **45**, 285-316(1990).

(64) Balzer, W.K., "Biases in the recording of performance-related information: the effects of initial impression and centrality of the appraisal task," *ibid*., **37**, 329-347(1986).

(65) Murphy, K.R., Garcia, M., Kerkar, S. et al., "Relationship between observational accuracy and accuracy in evaluating performance," *J. Appl. Psychol*., **67**, 320-325(1982).

(66) DeNisi, A.S., Robbins, T. and Cafferty, T.P., "Organization of information used for performance appraisals: Role of diary-keeping," *ibid*., **74**, 124-129(1989).

(67) Bernadin, H.J. and Walter, C.S., "Effects of rater training and diary keeping on psychometric error in ratings," *ibid*., **62**, 64-69(1977).

(68) Larson, J.R., "Role of memory in the performance evaluation process: with special reference to diary-keeping," *Psychol. Rep*., **57**, 775-782(1985).

(69) Guion, R.M., Personnel Testing, McGraw-Hill, New York NY (1965) pp. 466-475.

(70) Srull, T.K. and Brand, J.F., "Memory for information about persons: The effect of encoding operations on subsequent retrieval," *J. Verbal Learn. Verbal Behav*., **22**, 219-30(1983).

(71) Cafferty, T.P., DeNisi, A.S. and Williams, K.J., "Search and retrieval patterns for performance information: Effects on evaluations of multiple targets," *J. Pers. Soc. Psvchol*., **50**, 676-83(1986).

(72) Williams, K.J., DeNisi, A.S., Meglino, B.M. and Cafferty, T.P., "Initial decisions and subsequent performance ratings," *J. Appl. Psvchol*., **71**, 189-95(1986).

(73) Srinivas, S. and Motowidlo, S.J., "Effects of raters' stress on the dispersion and favorability of performance ratings," *ibid*., **72**, 247-251(1987).

(74) Borman, W.C., "Exploring upper limits of reliability and validity in job performance ratings," *ibid*., **63**, 135-144(1978).

(75) Nathan, B.R. and Lord, R.G., "Cognitive categorization and dimensional schemata: A process approach to the study of halo in performance ratings," *ibid*., **68**, 102-114(1983).

(76) Green, R.L., "Sources of recency effects in free recall," *Psychol. Bull*., **99**, 221-228(1986).

(77) Kaplan, M.F., "Stimulus inconsistency and response dispositions in formins judgments of other persons." *J. Personal Soc. Psychol*., **22**, 58-64(1973).

(78) Steiner, D.D. and Rain, J.S., "Immediate and delayed primacy and recency effects in performance evaluation." *J Appl. Psychol*., **74**, 136—142(1989).

(79) Kravitz, D.A. and Balzer, W.K., "Context effects in performance appraisal: A methodological critique and empirical study." *ibid*., **77**, 24-31(1992).

(80) Schmitt, N., "Social and situational determinants of interview decisions: Implications for the employment interview," *Personnel Psychol*., **29**, 79-101(1976).

(81) Regan, D.T., Straus, E. and Fazio, R., "Liking and the attribution process," *J. Exp. Social Psychol*., **10**, 385-397(1974).

(82) Cardy, R.L. and Dobbins, G.H., "Affect and appraisal accuracy; Liking as an integral dimension in evaluating performance." *J. Appl. Psychol*., **71**, 672-678(1986).

(83) Zalesny, M.D. and Highhouse, S., "Accuracy in performance evaluations," *Organizational Behavior and Human Decision Processes*, **51**, 22-50(1992).

(84) DeCotiis, T.A. and Petit, A., "The true performance appraisal process: a model and some testable hypotheses," *Acad. Management Rev*., **21**, 635-646(1978).

(85) Scheetz, L.J., "Measuring clinical competence in baccalaureate nursing students," in Measurement of Nursing Outcomes Volume Three: Measuring Clinical Skills and Professional Development in Education and Practice, (edits. Waltz, C.F. and Strickland, O.L.) Springer Publishing Company, New York NY (1990) pp. 3-15.

(86) Rossel, C.L. and Kakta B.A., "Clinical evaluation of nursing students: A criterion-referenced approach to clinical evaluation based on terminal characteristics." in *Measurment of Nursing Outcomes Volume Three: Measuring Clinical Skills and Professional Development in Education and Practice*, (edits. Waltz, C.F. and Strickland, O.L.) Springer Publishing Company, New York NY (1990) pp. 17-29.

(87) Howard, E.P., "Measurement of student clinical performance." in *Measurement of Nursing Outcomes Volume Three: Measuring Clinical Skills and Professional Development in Education and Practice*, (edits. Waltz, and Strickland, O.L.) Springer Publishing Company, New York NY (1990) pp. 31-43.

(88) Littlefield, J.H., "Developing and maintaining a resident rating system," in. *How to Evaluate Residents*, (edits. Lloyd, J.S. and Langlsey, D.G.) American Board of Medical Specialties, Chicago IL (1986) pp. 117-130.

APPENDIX A. DEFINITIONS OF COMMON PERFORMANCE APPRAISAL TERMS

Accuracy. A quality of a measurement that infers it measures the true performance. Accuracy implies the measures are valid and reliable, but the reverse is not true (53).

Assessment. A determination or appraisal of an individual's performance.

Attribution Theory. An interpretation of a behavior in terms of its causes.

Automatic Processing. The mental processing of information that requires no conscious effort and does not interfere with ongoing mental activities (45). The rater is unaware of this processing.

Central Tendency Error. A rating error that occurs because the rater hesitates to give extreme judgments such as the highest and lowest scores. All ratings tend to be near the mean score.

Clinical Evaluation System. The assessment methods and procedures established for faculty-student communication about student performance.

Cognitive Categorization. One of two theories explaining how a rater integrates multiple observations into a single rating. Cognitive categorization assumes behaviors are continually integrated into a general impression that serves as the basis for subsequent ratings(75).

Cognitive Processing. The mental processing of information related to performance appraisal. The steps involved in this processing are the acquisition, encoding, retrieval, and integration of performance information.

Dimensions. The independent qualities or components of the overall performance. For example, several components of competence in pharmacy practice are communication skills, patient monitoring skills and professionalism.

Encoding. The process of converting what is observed into some representation such as memory storage or written statements(56).

Evaluation. An inference of evaluation presumes the evaluation method can differentiate good from poor performance.

Feedback. An appraisal of performance that is conveyed between the instructor and the student during the learning period. This performance information is communicated to the student during a formative evaluation.

Extrapolation. An inference of validity that indicates the assessment method accurately predicts the student's performance in the actual practice setting.

Generalization. An inference of validity that indicates the assessment method can correctly predict performance in a variety of situations based on only a sample of test items or observations during the evaluation period.

Halo Effect. Rather than distinguishing among the levels of performance on different dimensions, the rater assigns the ratings based on a global impression.

Halo Error. The tendency for a rating to be influenced by traits or loosely associated factors. For example, the rater's scores are influenced by the student's attractiveness or personality.

Instrument. The tool that is used to assess a person's performance. For example, a rating form and a pencil-paper test are each instruments.

Interrater Reliability. The consistency of ratings between two or more raters who observe the same event and independently evaluate the performance.

Observation-based Ratings. An evaluation technique that requires the instructor to directly observe the student perform in the practice setting.

Rater Accuracy Training (RAT). A type of rater training that focuses on defining the dimensions comprising the rating instrument, and how to use the rating instrument.

Rater Error Training (RET). A type of rater training that emphasizes avoidance of the common rating errors such as halo, leniency, stringency, and central tendency.

Recency Effect. When integrating several observations of performance, the student's most recent observations have a greater effect on the overall rating.

Reliability. The degree to which ratings or a test scores are free from errors of measurement.

Stimulus-Based Judgments. One of two theories of how a rater derives a rating based on a series of observations. This theory-presumes the rater evaluates each individual incident during the observation period and then uses a weighting scheme to average them at the time of assigning a rating.

Stringency. The tendency of a rater to rate all students low.

Validity. The appropriateness, meaningfulness, and usefulness of the inferences made from ratings or test scores.