Criterion-Referenced Definitions for Rating Scales in Clinical Evaluation

Kathleen Nowak Bondy, RN, PhD

ABSTRACT

The evaluation of clinical competence of nursing students is often considered subjective and inconsistent. To avoid these pitfalls, the author developed criteria for a five-point rating scale for evaluation of student clinical performance. This rating system may be applied to any professional behavior and provides the student with diagnostic feedback as well as a fair assessment of performance.

Nursing, like many other academic disciplines, has a "hands-on" or clinical component which is integral to its educational progress. Historically and unlike most areas of study, however, students of nursing are expected to demonstrate competency in the clinical area as a condition for graduation. Even a student who receives A's and B's on written tests usually does not complete a nursing program if clinical evaluations show evidence of unsatisfactory performance.

The importance of clinical competence and faculty desire to be "objective" when evaluating a student's performance has led to much work by faculty members on clinical evaluation but with little research based upon psychometric principles. While some schools have developed minimal competency or critical element systems or clinical performance examinations in order to promote objectivity and fairness, most schools use some type of rating form. The purpose of this paper is twofold: 1) to explain a set of labels and their definitions which were developed for use with five-point rating scales for clinical evaluation; and 2) to present a framework for their use.

Background

Rating forms when used as clinical evaluation instruments consist of three parts: first, the clinical behaviors to be evaluated; second, a predetermined range of reference points on the rating scale; and third, a label for each reference or scale point. The clinical behaviors vary to reflect differences in the type of educational program, philosophy and mission of the school, the course objectives and content, and the level of the learner. In addition, clinical behaviors are developed to reflect psychomotor learning (as in skill development), affective learning (as in professional socialization or development), and cognitive learning (as in use of a knowledge base). On the rating form behaviors can be organized as a series of discrete statements or, as shown in Figure 1, as subsets headed by a general course objective and explicated by a number of specific behaviors. Achievement in a specified period of time such as a semester or a quarter is inferred by a composite weighted score of the behaviors over multiple observation points. Since much thought, time and effort are appropriately spent in determining and wording behaviors so that they clearly reflect many considerations, it is frustrating to faculty to see their effectiveness reduced by inadequate definitions, descriptions, or criteria for the reference points.

The reference points on a rating scale indicate the degree of competency or accomplishment with which the student has performed the behavior. Unlike checklists which utilize two-point scales (pass/fail, present/absent, satisfactory/unsatisfactory, and credit/no credit), rating scales can use from three to 20 scale points, and thus convey far more information about the quality of the performance and discriminate more accurately between learners. Psychometric studies indicate that the reliability of individual rating scales is a monotonically increasing function of the number of scale points; this reliability rises rapidly as the scale points increase from two to seven with little additional gain from 11 to 20. While it might seem that the reliability would decrease with additional scale points on a retest, the true score variance increases at an even more rapid rate than the error variance (Nunnally, 1978). Considering this, undefined or vague definitions for the scale points and rater or observer problems are more likely to contribute to low reliabilities than the number of scale points (DeMers, 1978, pp. 89-115).

Most rating scales in nursing have used five-point scales. It might be helpful to briefly review some of the scale
descriptors that are commonly used. The most common are abstract labels such as A, B, C, D, E or 5, 4, 3, 2, 1. What is the real difference between an A and a B or between a 4 and a 3? Is an A equivalent to a 5 or to a 1? Rarely are definitional statements available to faculty and students to accompany this type of abstract labeling.

In a similar manner, qualitative labels such as excellent, very good, good, fair, and poor usually lack a description of the criteria that make a performance excellent instead of good. Also, it is difficult to discern if a performance is excellent for that student, for that group, for all students, or for the number of students to which the instructor has been exposed. Normative labels such as superior, above average, average, weak, and failing raise similar interpretative problems. Even when a description of satisfactory performance is written into each behavioral statement, it is usually not sufficient guidance to discriminate between levels of performance and yet to write a full description of each behavior at each level is time-consuming and yields a very lengthy evaluation tool.

The problem is more acute with frequency labels such as always, usually, frequently, sometimes, and never. These pose problems when a student performs a behavior only once or the opportunity to demonstrate a behavior occurred
only once early in the clinical experience. On occasion, these terms more adequately describe the clinical site instead of the clinical student.

In summary, while all the label sets fit five-point scales, they are not necessarily applicable to all types of behavior or situations. When weighted to obtain a score (for possible translation into a grade), the scores may not necessarily mean the same thing. For example, if one instructor used frequency labels, another normative labels, and a third qualitative labels and all were weighted 5 (high) to 1 (low), a score of 3.5 on the behavior, able to make an unoccupied bed, may not indicate similar competency in that behavior even within the same educational setting. Therefore, even very conscientious and experienced faculty with carefully developed behavioral outcomes can experience qualms about the validity and reliability of clinical evaluation when definitions or criteria for the scale points are not explicit and explicable.

Given the situation, one would expect more confusion than has been experienced, but in fact, given a behavioral statement, clinical situations and level of learner, experienced faculty members through discussion can agree on the level of competency.

If areas of agreement evolve through substantial experience in clinical teaching, then identification of those elements and their development into a criterion-referenced set of scale labels may contribute to the validity and reliability of clinical evaluation.

Development of the Criteria

The criterion-referenced scale labels evolved from numerous discussions with clinical faculty, a comparison of the process of student learning with that of rehabilitation clients and literature on the process of skill acquisition (Harrow, 1972; Simpson, 1966). Two groups of experienced faculty judged preliminary drafts and suggested items which made the criteria useful for affective, cognitive and action-pattern (Simpson, 1966) types of behaviors as well as psychomotor behaviors. Comments from students and faculty when the criteria were used at several curriculum levels and in diverse clinical settings further refined the definitions. Since their conception, over 150 faculty have reviewed and commented favorably on the validity, comprehensiveness and manageability of the criterion-referenced scale definitions. Additionally, literature appearing subsequent to their development validates the included criteria (Crocker, 1976; Smania, McClelland, and McClosky, 1978).

Explanation of the Criteria

The characteristics of competency or criteria for clinical evaluation cluster into three major areas:

1. Professional standards and procedures for the behavior.
2. Qualitative aspects of the performance.
3. Assistance needed to perform the behavior.

Five levels of competency are identified which can be descriptively labeled: Independent, Supervised, Assisted, Marginal, and Dependent, and which are applicable to each of the three major areas. The label names were chosen for their overall descriptive value and differences from other forms of labels. In this context, Independent means meeting the criteria identified in each of the three areas; it does not mean without observation, for the performance must be observed to be rated independent by someone other than the performer or client. A student can demonstrate independent judgment when appropriately requesting assistance or guidance in a situation. In similar manner, each of the levels are defined by the description of the characteristics in the three areas, as shown in Figure 2.

PROFESSIONAL STANDARDS: The intent of educational programs in nursing is to produce practitioners who have acquired a knowledge base, therapeutic and interpersonal skills, and values and attitudes that characterize the nursing profession, are safe for the public and reflect the philosophy of the school. Standards are set by the profession and transmitted through the professional educational process (Smania et al., 1978).

Professional standards, when interpreted in terms of safety, accuracy, effect, and affect, can be applied to behaviors in three domains: cognitive, affective and psychomotor. A behavior should be safe not only for the client but for the nurse and others in the environment as well. Contaminating an area normally considered clean in strict isolation without correcting the situation makes that environment unsafe for others. Accuracy and precision can be evaluated in applying the knowledge base to the clinical situation, in verbal, non-verbal and written communications, in the appropriate use of one’s professional vocabulary, in approaches to various situations as well as in commonly recognized psychomotor skills.

EFFECT: Effect refers to achieving the intended purpose of the behavior. For example, if the desired effect of interviewing a client is to assess the current health status, the evaluation of performance is in part based on whether appropriate information was requested and obtained; i.e., if the purpose of a bath is to cleanse a client’s body, is that body now clean?

AFFECT: Affect refers to the manner in which the behavior is performed and the demeanor of the student. This could influence the effect of the intervention on a client. For example, when a nurse enters a room to turn a spinal cord-injured client who has begun to question his life’s value, it could make a considerable difference to his self-concept whether the nurse’s tone of voice and demeanor convey that this is important because he is valued as a person or that this is just another boring task to be performed.

In describing the professional standards and procedures area at five levels, the performance must be safe, accurate
with the appropriate effect and affect each time at both the Independent and Supervised levels. At the Assisted level, the effect and affect are appropriate most of the time. At the Marginal level, the student performs with risk to the client, the student or others or the student may be safe but not when alone; usually the latter is inferred from unsafe or at risk behavior on previous occasions, questionable safety as reported by others or the student’s ability to describe or discuss the intended behavior raises questions of safety. At this level, the student is not always accurate, and occasionally achieves the accepted level or affect. At the Dependent level, the student’s performance is unsafe or the student cannot demonstrate the behavior.

QUALITY OF THE PERFORMANCE: The second area incorporates qualitative aspects of the performance. These are based upon degrees of skill development which encompass the use of time, space, equipment, and the utilization or expenditure of energy. While the total range of skill development extends from mastery or expertness through an inability to do something, the range in these criteria begins with proficiency in basic programs. While students could have “mastered” selected skills, basic programs prepare beginning practitioners, not master practitioners.

At the Independent level, proficiency means usual efficiency and implies exceptional deftness, use of subtle perceptual cues to modify the behavior to achieve the desired effect and exceptional coordination and integration. The sequence of movements and communication are fluid, even and intertwined. There is an economical use of movements, equipment, and conversation. The behavior is
demonstrated within an expedient or minimal time period as when giving a STAT medication. The student appears confident and relaxed and only occasionally or subtly expends excess energy in performance. This level of performance is seen when the student focuses on the client rather than on the skill that is being performed.

At the Supervised level, the student is efficient and coordinated but expends more of her energy or that of the client in accomplishing the behavior. She appears confident and focuses on the client but can be distracted to the skill as it becomes more complex. The behavior is performed within a reasonable time period; for example, she can prepare and administer a routine IM injection within an acceptable time period from the specified time but becomes flustered under a STAT situation.

At the Assisted level, the student is skillful in parts of the behavior while the rest of the performance is characterized by inefficiency and incoordination, thereby expending excess energy in movements, as in selecting inappropriate supplies in type or number. At times the student appears anxious, worried or flustered, but makes an effort to appear confident. Accomplishment of the behavior takes longer and the end result is sometimes late. The student focuses more attention on the behavior or on herself than on the client.

At the Marginal level, the student is unskilled, inefficient and expends considerable excess energy in performance; little thought appears to have been given to the sequence of activities that is to be performed. Anxiety may be apparent or masked. Completion of the behavior is considerably delayed to the extent that other activities are disrupted or omitted.

At the Dependent level, the student may attempt the procedure but is unsuccessful; unreasonable energy may be expended in attempting the procedure or the student appears unable to move.

**ASSISTANCE REQUIRED:** The third area that contributes to evaluation is the type and amount of instructor assistance or cues needed to demonstrate the behavior (Smania et al., 1978). Cues can be supportive or directive. Cues such as “that’s right,” “keep going” and the like are supportive, encouraging or reinforcing but do not change or direct what the student does or says. Directive cues, which can be verbal and/or physical, indicate either what to do or say next or correct an ongoing activity. Reminding a student to check a nameband as she hands the medication to the client is a directive cue; making the same comment before entering the room is superfluous since it interferes with observing what the student knows or will do. Telling a student she did well after the performance is not a cue but feedback information; cues refer to what is necessary to maintain or encourage the student’s performance. Frequently, the cues a student needs can indicate that student’s ability to follow directions or her anxiety level.

At the Independent level, a behavior is performed primarily without supporting cues or the student did not need the cues that were given. At the Supervised level, occasional supporting cues and infrequently a directive cue are needed. At the Assisted level the student requires frequent verbal and occasional physical directive cues in addition to supportive cues; at the Marginal level, continuous verbal cues and frequent physical cues are required; at the Dependent level, the verbal and physical cues are so directive and continuous that essentially it is the instructor who actually performed the behavior.

These three basic areas of criteria are usually observed simultaneously and are taken together to determine the level of a student’s performance on a given behavior. The behavior is evaluated at the lowest level of achievement in any of the three areas. For example, if the behavior is the ability to give a complete bed bath and the student performed it correctly (independent), efficiently within a reasonable time frame (supervised), but required frequent verbal in addition to some physical directive cues (assisted), then that performance is evaluated at the assisted level. The issue of whether the performance by standards is independent or supervised is moot since the lowest level determines the overall evaluation of the behavior. This, then, reflects the minimal level of performance which can be expected at another time or in a similar situation and most closely approximates reality. A focus on the best parts of the performance occludes those areas which would benefit from teaching and which may interfere with an expected performance level at another time or in a similar situation. The criteria by which the student demonstrated greater competency is more reflective of potential performance than actual performance when reported. In an effort to give students positive support and reduce anxiety, there is a tendency to overrate the strong aspects of a performance and underrate the weak aspects of the same performance.

If a part of a behavior is omitted, then the instructor must judge whether the omission relates to the acceptability or to the quality of the task. For example, failure to check a nameband when giving a medication pertains to the acceptability of the performance, but failure to ask if the client is comfortable or needs anything before leaving the room pertains to the quality of a performance. An X or NO (not observed) column is employed to protect the student from lack of opportunity to demonstrate a behavior and/or the necessity of random sampling of behavior. Use of the criteria assumes that the behaviors used are appropriate for the level of the learner and generally available in the clinical area. Therefore, a category of “not appropriate” is not necessary in addition to a “not observed” column.

Up to this point the discussion of the criteria has focused on observing a performance of a single behavior. In the clinical experience, many observations are made and these need to be summarized in some manner. Over the clinical experience, students tend to demonstrate patterns of development (Figure 3). The expected pattern of student performance is that when presented with new or more complex behaviors, students’ level of competency is lower and rises as they learn to demonstrate the expected behavior. There is an unstated assumption by most instructors that a student’s performance will be generally consis-
tent with the expected Pattern I or similar to Patterns II and IV which show some degree of predictability. Consistency and predictability of performance are aspects of competence among professionals so hence, an erratic series of performances as in Pattern V is professionally considered provisional or even dependent. A disintegrating performance as in Pattern III may be the result of numerous causes. However, if appropriate counseling and problem solving have been done, a sterling performance on the last clinical day more accurately reflects potential rather than reality; the consistency of performance is a more critical criterion for evaluation.

The use of pattern of performance in deriving a summative evaluation implies that I, S, A, M and D cannot be numerically weighted, summed over the semester and then averaged for a score. For example, if the weighting were 5, 4, 3, 2 and 1, respectively, the student in Pattern I with steady development would achieve a score no higher than 2.0-3.0 which does not reflect the correct competency level at the end of the course for that behavior.

### Grading

The summative levels of competency over behaviors (or objectives) can be weighted, summed and averaged when an overall value is necessary for grading purposes.

Assuming a weighting of 5, 4, 3, 2, 1 for I, S, A, M, D, respectively, acceptable minimal clinical scores are set by a faculty referent to the definitions these numbers represent. One faculty group may set 3.0 as the minimal overall score and another group may use 2.75 or 3.25. The clinical score can be translated to other equivalent forms as letters, percentages or points as required. The equivalencies are also determined by the faculty using the criteria and, perhaps in part, by policies of the school. For example, a clinical score of 3.0 could be a C, C− or even B−; it could be 70%, 75% or 80%. The point is that these are faculty decisions which may vary from one school/course to another. The choice of letters, percentages or other forms may depend on whether the score stands alone or is incorporated with other material for a course grade. In other words, a “grade” is extrinsic to the criteria; the criteria make a grade more interpretable. The role of the clinical instructor primarily is to compare a student’s performance to the criteria; the assignment of “grades” is relative to the criteria, not to the individual student.

### Implications

Use of the criteria has several benefits both for individual faculty and students as well as collectively within a program. Individual faculty members who have used the criterion-referenced scale definitions comment that it enables them to describe and classify more accurately the strengths and limitations of a student’s performance. Consequently, students are given more diagnostic feedback and perceive the comments as constructive and positive rather than as critical or negative in nature. Students have stated that the criteria “make sense” to them and that evaluation of performance is more “above board” and reduces perceived subjectivity.

When the criteria are the common base for discussion between student and instructor, the student learns to self-evaluate and to validate self-perceptions of performance.
Several faculty have commented that the criteria can be actively used to teach students to self-evaluate with validation from the instructor. An occasional student has observed that the same words can be used to describe a client's progress in learning self-care skills.

In discussing student performance in general or in particular with colleagues, faculty have stated that they begin to identify various strategies that are more likely to succeed in assisting a student to improve performance. A student who performs the behavior with excessive energy and time requires a different approach to improve than either one who does not know how to perform the behavior or one who is inordinately relying on supportive cues. Using the criteria as a common framework within which to discuss student performance, faculty members are better able to identify the norm behaviors and competency levels achieved by students at the individual curricular levels. Thus, articulation between clinical courses and leveling of behaviors is enhanced.

The use of the criteria schema in all clinical courses decreases students' need to "psych out" the implicit criteria used by individual instructors and to use that energy instead to understand the expected behaviors and/or raise their competency levels; in other words, to learn. In like manner, a clinical evaluation in one clinical course can have some relationship to a clinical evaluation in another course.

In addition, the criteria can be used to give new faculty or clinical instructors a standardized framework within which to observe student clinical behavior.

Conclusion

When carefully developed, rating scales in clinical evaluation of student performance can be psychometrically effective and efficient. This includes equal attention to both the clinical behaviors and criteria for competency. Undefined or vague labels for the scale points can contribute to lower reliability. These criteria and their scale point descriptors have, in pilot testing and use over seven years, demonstrated their content validity and utility. Research is in progress to test the extent to which they affect the reliability of clinical evaluations of student performance.

References


Nunnally (1978).

