ANALYSIS OF STUDENT PERFORMANCE RATINGS

MONICA D. ANGUS, PhD.

Evaluation procedures in schools of nursing differ depending upon tradition and the orientation of those responsible for student assessment. Traditionally, faculty in schools have used anecdotal or graphic ratings which attempt to give an overall estimate of the student's capability. While these global estimates are useful in assessing whether or not the student meets some minimum standard of "safety to practice" they provide little information about differences in competencies between students or classes in a program.

Proponents of the more recently developed behavioral rating scales (sometimes called "BES," sometimes "BARS" or "BOS") express the need to assess performance in terms of behaviors which are critical to job success or failure (Campbell, Dunnette, Lawler and Weick, 1970). The reason for this is that behavioral measures are based on what a person actually does as opposed to what might be inferred from factors which are not entirely under his/her control or from attitudes or traits (Latham and Wexley, 1977).

If performance assessment is based on relevant behaviors, it can provide information which will assist students with their professional development. Further, if similar measuring instruments are used in more than one school and on more than one class in each school, they can provide valuable information to program planners both within and across settings. The following study reports the results of using a particular variant (BES) of behaviorally anchored ratings to assess student nursing performance in two diploma programs. Outcomes in the study will focus on identifying information relative both to student and program development.

SCALE PREPARATION

There are two stages in the development of Behavioral Expectations Scales (BES). The first is termed scale development and the second, scale operation. The procedures for scale development first outlined by Smith and Kendall (1963) utilize the Critical Incident technique of Flanagan (1954). As it is not altogether clear that each step in the lengthy development process has to be repeated by every user group in order that the scales possess good psychometric properties, researchers have often modified the technique. As the focus of this study is placed on the results of scale operation, the original procedures for scale development were, here, slightly modified.

A. *Scale development*

The first step in scale development is to have experts in the field determine what areas of performance they feel are important to measure. Emphasis is placed on defining areas which are discriminable behaviorally. Once agreement is reached on definitions for these areas, items (critical incidents) are written to measure the areas. The procedures next focus on determining the level of performance (scaling) each item illustrates. Item statistics are calculated and the final format assembled.

The rating scales for this study were developed by faculty of two schools of nursing with the assistance of the author. Faculty reviewed the definitions of five rating scales used in previous research with graduate nurses (Tate, 1964) and agreed that definitions of the five areas (Knowledge and Judgment, Conscientiousness, Skill in Human Relations, Organizational Ability, and Observational Ability) were acceptable. Further, they agreed that these areas are important ones which need to be measured in an assessment of performance and assigned equal weight to the five.

As a second step, faculty assigned a set of written items (nursing behaviors) compiled from several sources (Slater Nursing Competencies Rating Scale, 1967; Quality Patient Care Scale, 1970; and the original scales developed by Tate, 1964) to the one of the five areas the item best described. When making the assignment, individuals were asked to determine first what area the item best described and then to determine at what level of performance they would rate the behavior (scaling). Since several behaviors described graduate rather than student nurse behavior, raters were asked to attend particularly to those items which had relevance for *student* performance. Faculty worked independently. If seventy-five percent or more of the faculty agreed on the assignment of an item, it was allocated to that scale. Then, if the standard deviation among raters' assignments to a scale level was not greater than .75, the item was retained to become an anchor on one of the five scales. Figure 1 is an example of an anchored scale.

Figure 1

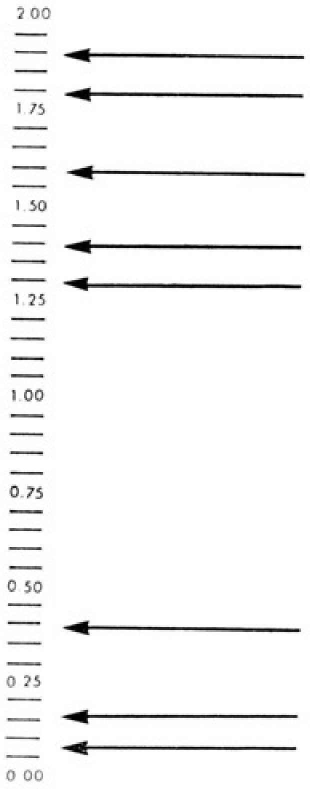
SKILL IN HUMAN RELATIONSHIPS (WITH PATIENTS, FAMILIES AND CO-WORKERS) – **Behaves in a manner appropriate to the situation and individuals involved**

NOTE: From Nursing Performance Rating Scale by L.M. Rendall and M.D. Angus. 1974. Reprinted by permission.

Even when there is considerable emotional self-involvement, behavior with others is so skillful and insightful that it not only smooths but often prevents difficult emotional and social situations; this implies the ability to recognize the subtle as well as the more obvious components of basic emotional reactions in self and others (e.g. anxiety, fear, frustration, anger, etc.).

If emotional self-involvement is minimal, behavior with others is such that it does not complicate difficult emotional and social situations; this implies the ability to recognize the more obvious components of basic emotional reactions in self and others.

Behavior with others is such that it tends to complicate or create difficult emotional or social situations; this implies an inability to recognize even the obvious basic emo...onal reactions of self and others.



This student could be expected, whenever possible, to sit down and talk with a terminal cancer patient who is considered to be "demanding".

This student could be expected to encourage a terminal patient to talk of death and dying.

This student would help a patient to understand the appropriateness of dependent role and assure him that he will be helped to resume independence as soon as possible.

If a patient assumed a threatening posture, this student would speak quietly to him.

This student could be expected to arrange to have the patient's discussion with allied professionals conducted in privacy.

If a convalescent patient complained about the service, this student would be likely to tell the patient that there is a shortage of nurses and the needs of the sickest patient have to be met first.

In the presence of a woman who is crying because her husband is dangerously ill, this student would be expected to tell the woman not to cry.

If this student were told by an ambulatory patient that a patient in the ward was having difficulty in breathing, could be expected to tell the patient that his help was not needed.

When the scales were assembled, faculty in a two year diploma program used the scales to evaluate student performance. These faculty members had participated in scale development. A small number of faculty from a three year hospital based diploma program also used the scales. These people had not participated in scale development. The following report deals with the results of these faculty ratings.

B. *Scale operation*

A total of 109 student ratings were obtained for this study. Faculty in the two year program rated students from two classes: thirty-five students in one class and thirty-one in the other. In the three year program, ratings were performed on fifteen students. All students were evaluated at the end of the first year in the program. In addition, students in the second class in the two year program were evaluated a second time just before graduation. (A few students were not available for this assessment because they had left the program.)

Statistical analyses of the data were performed using the programs "Corr" and "Oneway" ANOVA from the statistical package for the Social Sciences. The BMD12V program for multivariate analysis of variance from the Biomedical Computer programs was also used.

RESULTS

The means and standard deviations for ratings on the separate scales are presented in Table 1. Table 2 presents the correlations between the five areas of behavior rated.

The first thing to note with respect to Table 1 is that mean values over ratings are highest on the measure of Skill in Human Relations and Organizational Ability followed by Conscientiousness, Observational Ability, and Knowledge and Judgment. Since the scales range in value from 0.00 to 2.00, it is clear that the average scale ratings are all above average for this group of ratings. On the whole, differences in mean values are small.

TABLE 1
SCALE MEANS AND STANDARD DEVIATIONS. $N=106$

Scale	Mean	Standard Deviation
Knowledge and Judgment	1.21	.36
Conscientiousness	1.29	.36
Skill in Human Relations	1.32	.37
Organizational Ability	1.31	.37
Observational Ability	1.29	.37

TABLE 2

CORRELATIONS BETWEEN THE FIVE SCALES

Scale	1	2	3	4	5
1. Knowledge and Judgment	—				
2. Conscientiousness	.64*	—			
3. Skill in Human Relations	.59*	.67*	—		
4. Organizational Ability	.50*	.68*	.43*	—	
5. Observational Ability	.67*	.77*	.65*	.69*	—

*$p<.001$

Table 2 illustrates the correlations between the five areas of student nursing behavior rated. They appear to be moderately to well correlated. The largest correlation is between Observational Ability and Conscientiousness ($r=.77$) and the lowest between Organizational Ability and Skill in Human Relations ($r=.43$). It is interesting to note that for the last two scales the group mean is highest and the correlation lowest. It suggests that to be an effective communicator is not necessarily to be a good organizer. The correlations between other scales lie somewhere in the range between .43 and .77. All are significantly different from zero.

It would appear that as a general rule, if a student is rated high in one area, she will be rated high in another and vice versa. A visual inspection of plots of data for individual students illustrates that this is so. For example, one student was consistently rated near the top of each scale. Her ratings are: 1.90, 1.90, 1.76, 2.00 and 1.78. Since she has reached, or is near, the ceiling on almost all scales, consideration should be given to advancing her in the program or providing her with an enriched program.

Another student, however, was consistently rated below average on the five scales. Her ratings are: .25, .25, .25, 1.00 and .75 for the five scales respectively. It seems that this student is performing at levels well below that which most faculty would find acceptable. If faculty move to develop cut-off scores on the scales to indicate the level of performance they will accept, this student might well be failed.

In general, while it seems that most students' ratings are slightly above average, there are a number who are consistently rated below average. It is this group of students that faculty should be most concerned about with respect to "safety to practice."

Some student's ratings on the five scales are disparate. That is, a student may receive a high rating on one scale and a low rating on another. While such students are atypical, their ratings are important for they illustrate how the BES technique assists faculty to identify a student's strengths and weaknesses.

Figure 2 depicts graphically three student's mean ratings in the five areas measured. Student #1 received ratings of: .60, 1.30, .70, 1.60, and 1.60 respectively, rating below average on Knowledge and Judgment and Skill in Human Relations and above average on Conscientiousness, Organizational Ability and Observational Ability. This student has some of the qualities necessary for good nursing performance — that is, she possesses above average organizational and observational ability — but she appears to be so lacking in Knowledge and Judgment and Skill in Human Relations as to be "unsafe to practice." If she is to successfully complete the program, she will need to apply herself more to her studies and learn to communicate better.

Student #2's scores are: .75, 1.50, .30, 1.60, and 1.05 While this student scores well above average in Conscientiousness and Organizational Ability, her ratings in the area of Knowledge and Judgment indicate she needs to be counseled to improve her performance. However, her ratings in the area of Skill in Human Relations are so low it is clear she has great difficulty with her interpersonal relations. Perhaps she is unsuited to nursing in spite of her conscientious attitude toward her duties.

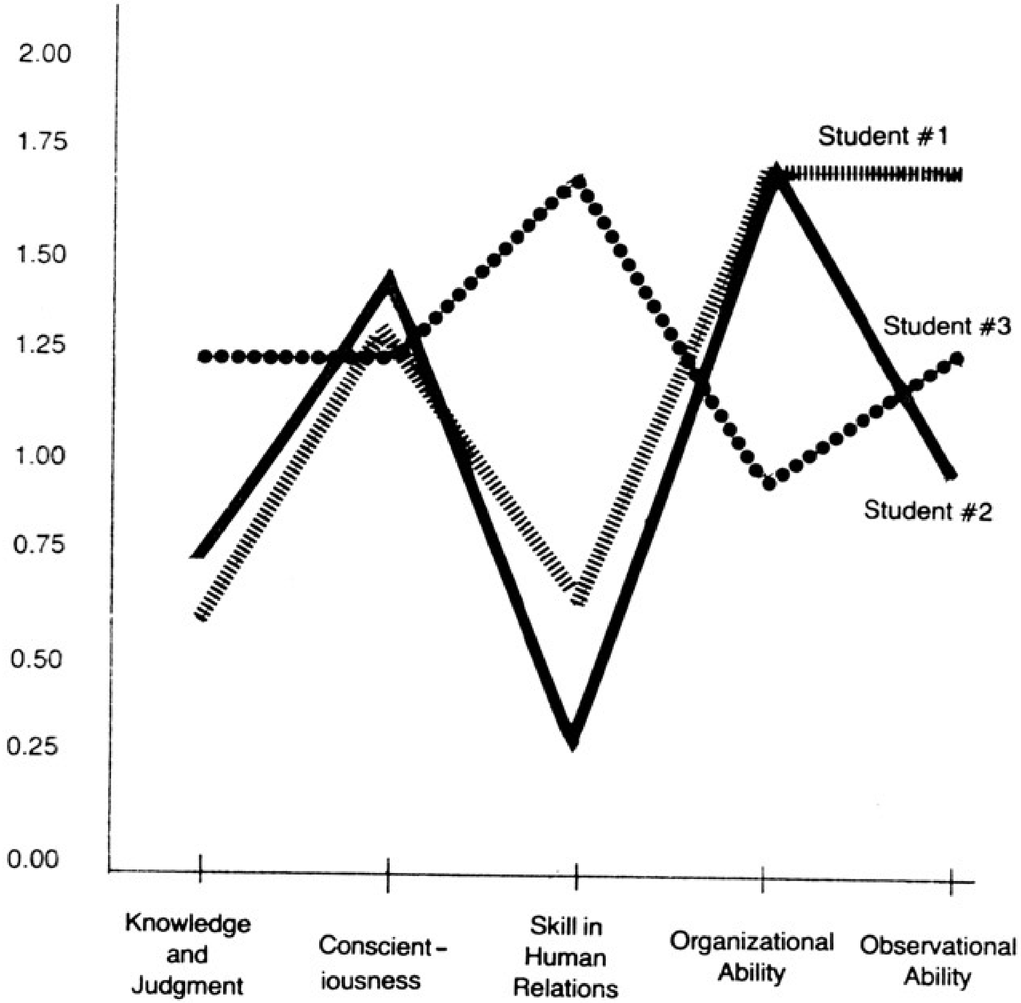
The scores of student #3 also illustrate the value of attempting to discriminate between student competencies in nursing. This student received ratings of 1.25, 1.25, 1.60, 1.00 and 1.25 on the five scales. While she is average or above on most qualities it is obvious that her strengths in nursing abilities lie in the area of Skill in Human Relations. She appears to be a very effective communicator. For this student, emphasis should be placed on improving her organizational skills. In reporting her evaluation to her, her strength in human relations should be brought out for positive reinforcement is something that will encourage her to repeat this behavior.

Ratings over time

Since there were ratings over time for twenty-eight students in Class 2 in the two year program, correlations between these students'

Figure 2

MEAN RATINGS OF THREE STUDENTS ON THE FIVE SCALES



ratings on the five scales were computed. The correlations were: Knowledge and Judgment (r=.65); Conscientiousness (r=.55); Skill in Human Relations (r=.54); Organizational Ability (r=.27); and Observational Ability (r=.37). All correlations except that for Organizational Ability are significantly different from zero ($p<.05$). It seems clear that except for the area of Knowledge and Judgment, ratings over time for these students are not highly related. This finding will be discussed later.

Group differences

In order to assess whether there are any important differences between classes of students in the two year program with respect to the qualities rated, a multivariate analysis of variance test (MANOVA) was performed on the data collected from evaluations done after a year in the program. Since the particular computer program used for this analysis requires equal *N*, data from four student's ratings were randomly deleted from Class 1. Results revealed a significant difference between classes ($F=7.9$ $df=5,60$ $p<.001$).

A series of Oneway ANOVA tests on the above data revealed a significant difference between the groups only on the Organizational Ability scale ($F=23.65$ $df=1,60$ $p<.001$) The mean of ratings for Class 1 is 1.31 and the mean for Class 2 is 1.02 on this scale. Organizational ability is present to a much greater extent in students in Class 1 at the end of the first year of the program.

A MANOVA test was performed on the data collected from students in Class 2 of the two year program both after a year in the program and just prior to graduation to see if there were any differences in abilities over time. Results revealed a significant difference between ratings ($F=15.96$ $df=5,60$ $p<.001$).

A series of Oneway ANOVA tests on the preceding data indicates a significant difference in all areas except Knowledge and Judgment (Table 3). While ratings in this area increased, the difference is not significant. The greatest improvement in scale ratings from one year to the next is in Organizational Ability. In this area the average scale rating increased from 1.02 to 1.62 — from average to well above average.

DISCUSSION

Effective performance appraisal systems are very difficult to build. Part of the reason for this is that different organizations have different philosophies of evaluation. On the one hand, most organizations use appraisals primarily to promote, transfer, release, and pay employees. Very often, the appraisal exists as a means to motivate employees to increased productivity — a tool to prod workers. In this instance, it often becomes a negative force and both supervisors and employees shun the task.

On the other hand, certain organizations and individuals within organizations feel that the appraisal should be used to assess personal development and that the promotion, transfer and release decisions are of secondary importance. Some employees, and certainly students, actively seek evaluation in order to plan for career development. In some institutions, particularly educational, supervisors assume that

TABLE 3

INDEXES DERIVED FROM ONEWAY *ANOVA* TESTS OF RATINGS ON THE FIVE PERFORMANCE SCALES OVER TIME. CLASS 2 $N=28$

Scale	*M*	*SD*	*F* ratio	*F* Prob.
Knowledge and Judgment				
Time 1	1.18	.41	.749	.39
Time 2	1.27	.44		
Conscientiousness				
Time 1	1.20	.40	14.01	.0004
Time 2	1.57	.38		
Skill in Human Relations				
Time 1	1.29	.39	3.55	.06
Time 2	1.49	.44		
Organizational Ability				
Time 1	1.02	.23	66.49	.000
Time 2	1.62	.33		
Observational Ability				
Time 1	1.21	.33	16.24	.002
Time 2	1.55	.32		

part of the job is to help those they supervise to increase their knowledge and skills. A primary objective of assessment then is to identify strengths and weaknesses in order to guide those supervised. In this instance, evaluation becomes a positive force for growth.

In nursing, those responsible for student assessment often experience role conflict. On the one hand, they are responsible for insuring that people who are not "safe to practice" are prevented from doing so and, on the other hand, they are responsible for the student's achieving proficiency. If incompetent students are not failed, there are serious implications for patients. However, such failures also have serious implications for students and faculty members. All of us who teach experience the conflict in assessments but surely it must be greater for those where human life is at stake. Therefore, the tools designed to assist faculty with student appraisal should be the best possible. It is suggested here that the BES technique is a method which provides rater and ratee with a maximum amount of information about behavioral performance. As a result, it is admirably suited to training situations.

While it is apparent that the competencies assessed is this study are moderately correlated, the evidence would not suggest that one could measure a single aspect of nursing behavior and achieve the same result as if all competencies are assessed. To be sure, as with global ratings, one could assess performance using a single scale and have a reasonably good chance of predicting "safety to practice." However, the margin of error would appear to be much greater when single rather than multiple scales are used to make judgments about student capability. When all five areas of nursing competency are assessed much more information is provided. Such information can assist faculty with both program planning and counseling individual students.

Of considerable interest in this study is the fact that ratings over time for students from one of the classes are not highly correlated. Just what is contributing to the variance here is not clear. Two explanations come to mind. The variance is due either to raters using different standards to rate performance (in spite of scales which have been developed to assist raters to avoid this) or students behaving inconsistently. A third explanation, of course, is that the variance is due to both of these factors. Further research is necessary to explore this finding.

The study provides some information for program planners about group differences. At the end of the first year of study, students in Class 1 of the two year program have higher ratings in Organizational Ability than do students in Class 2 of the same program: the ratings of Class 2 are only average. However, the ratings of Class 2 in the area of Organizational Ability improve over time; for this class, this area of competency showed the greatest improvement. However, more information is required to determine whether faculty placed more emphasis on training in this area or whether score variations were due merely to rater differences.

The higher ratings assigned to Class 2 for all assessed competencies at the end of the program is to be expected: one assumes that competency increases with time spent in the program. What is interesting is that ratings in the area of Knowledge and Judgment did not improve significantly. Findings reveal that faculty judge this area to be the weakest of all for students graduating from this program. Perhaps in the second year of a two year program faculty emphasize more clinical than cognitive skills.

A frequent complaint of those who have used BES to rate job performance is the amount of time it takes to do the job properly. For example, faculty in the two year program not only worked on scale development but participated in scale operation. They wrote down the behaviors they observed on the wards for each student assigned to them. Then they rated these observed behaviors using the developed scales. Thus faculty had no need to rely on memory or impressions. Reports from those in the two year program indicate the effort was worthwhile: faculty felt the technique helped them to study more carefully the behaviors of students.

However, faculty in the three year program where a few ratings were done for this study asked to be relieved of the task. Their spokesman suggested that the work schedule of faculty did not permit time to document each student's case in as thorough a manner as the BES technique requires. Instead, this group used a five point Likert type scale to rate performance. It may be that participation in the development of rating scales brings about a stronger commitment to their use (Borman and Vallon, 1974).

While this report focuses on the amount of information that can be obtained if the BES technique for evaluating performance is used, there is an important issue that has not been addressed in this paper. This issue relates to the superiority of BES with respect to other types of ratings. Results from studies which have put BES into competition with simpler formats are mixed, but on the whole, BES demonstrates psychometric superiority (Maas, 1965; Campbell, Dunnette, Arvey & Hellervik, 1973; and Burnaska and Hollman, 1974).

However, BES has not yet been put into competition with the Behavioral Observation Scales (BOS) developed by Latham and Wexley (1977). These investigators suggest that BES will likely be preferred when there is minimal opportunity for the manager to observe a subordinate and BOS when there is a high degree of contact between rater and ratee. They readily admit that the BOS type checklist requires raters to spend a greater amount of time rating.

In conclusion, although behavioral ratings appear to be superior to other types of assessment formats, there is some question as to which behaviorally anchored format is best. Schools of nursing should be active in conducting research to find out which approaches improve the quality of their assessments of nursing competence.

The intention here has been to report the results of using a particular behaviorally anchored technique to assess the performance of students in nursing. The amount and type of information the resultant ratings provided would appear to recommend their use.

REFERENCES

Borman, W. C. and Vallon, W. R. A view of what can happen when behavioral expectation scales are developed in one setting and used in another. *Journal of Applied Psychology,* 1974, *59,* 197-201.

Burnaska, R. F. and Hollman, T. C. An empirical comparison of the relative effects of rater response biases on three rating scale formats. *Journal of Applied Psychology,* 1974, *59,* 307-312.

Campbell, J. P., Dunnette, M. D., Arvey, R. D., and Hellervik, L. V. The development and evaluation of behaviorally based rating scales. *Journal of Applied Psychology,* 1973, *57,* 15-22.

Campbell, J. P., Dunnette, M. D., Lawler, E. E., and Weick, K. E. *Managerial Behavior, Performance and Effectiveness.* New York: McGraw-Hill, 1970.

Latham, G. P. and Wexley, K. N. Behavioral Observation scales for performances appraisal purposes. *Personnel Psychology,* 1977, *30,* 255-268.

Maas, J. B. Patterned scale expectation interview; reliability studies of a new technique. *Journal of Applied Psychology,* 1965, *49,* 431-433.

Schwab, D. P., Heneman, H. G., and De Cotiis, T. A. Behaviorally anchored rating scales; A review of the literature. *Personnel Psychology,* 1975, *28,* 549-562.

Smith, P.C. and Kendall, L. M. Retranslation of expectations; An approach to the construction of unambiguous anchors for rating scales. *Journal of Psychology,* 1963, *47,* 149-155.

Tate, B. L. *Test of a Nursing Performance Evaluation Instrument.* New York: National League for Nursing, 1964.

RESUME

ANALYSE DE L'EVALUATION DES RESULTATS D'ETUDIANTS EN MATIERE DE COMPETENCE CLINIQUE

On a utilisé une variante de la technique des attentes comportementales en sciences infirmières pour examiner l'évaluation de 109 étudiants en matière de compétence clinique. L'objet de l'analyse était de fournir des données aux professeurs sur chaque étudiant et sur les progrès de la classe durant le programme. Pour la majorité des étudiants, l'analyse a montré une corrélation moyenne quant au facteurs étudiés; pour quelques individus, les résultats sont divergents. On trouvera dans l'article le profil des résultats de certains étudiants. On a enregistré des différences significatives quant aux facteurs entres les différentes classes d'une même école et entre les années d'un programme pour une même classe. Il semble que ce soient les techniques d'évaluation offrant des données sur les comportements qui jouent un rôle critique dans la réussite ou l'échec professionnel qui seront le plus utiles aux enseignants pour l'orientation des étudiants et la planification des programmes.