Psychometric Characteristics of the Behavioral Observation Scale

Gregg Allen Bolt
Western Michigan University

Follow this and additional works at: https://scholarworks.wmich.edu/masters_theses
Part of the Industrial and Organizational Psychology Commons

Recommended Citation
https://scholarworks.wmich.edu/masters_theses/1482

This Masters Thesis-Open Access is brought to you for free and open access by the Graduate College at ScholarWorks at WMU. It has been accepted for inclusion in Master's Theses by an authorized administrator of ScholarWorks at WMU. For more information, please contact maira.bundza@wmich.edu.
PSYCHOMETRIC CHARACTERISTICS OF THE
BEHAVIORAL OBSERVATION SCALE

by

Gregg Allen Bolt

A Thesis
Submitted to the
Faculty of The Graduate College
in partial fulfillment of the
requirements for the
Degree of Master of Arts
Department of Psychology

Western Michigan University
Kalamazoo, Michigan
August 1984
PSYCHOMETRIC CHARACTERISTICS OF THE
BEHAVIORAL OBSERVATION SCALE

Gregg Allen Bolt, M. A.
Western Michigan University, 1984

Self-, peer-, and supervisor ratings were obtained on 52 psychiatric aides using a Behavioral Observation Scale (BOS). Self ratings showed less leniency error than peer- and supervisor ratings. Halo error could not be assessed due to a negative correlation between means and variances. A multitrait-multimethod (MTMM) analysis supported the presence of strong rater bias and significant convergent validity but not discriminant validity. The results of the analyses demonstrated that the ratings obtained from a BOS were not psychometrically superior to other appraisal formats. Questions were raised as to the adequacy of a five point scale, data transformation, and rating scales.
ACKNOWLEDGEMENTS

I would like to express sincere gratitude to the many people involved in the writing of this thesis.

Thanks goes to Gerald DeWeerd and the nursing supervisors who were willing to undertake this research project. Norman Peterson deserves special thanks for his undying willingness to advise me to and from Grand Rapids, and for his words of encouragement. A special thanks also goes to Peninnah Miller and Bradley Huitema who provided statistical assistance and consultation. I would also like to acknowledge Dale Brethower and Jack Asher who served as committee members. And to Sheryl, my wife, a special thanks for providing encouragement and support when the obstacles seemed insurmountable.

Gregg Allen Bolt
INFORMATION TO USERS

This reproduction was made from a copy of a document sent to us for microfilming. While the most advanced technology has been used to photograph and reproduce this document, the quality of the reproduction is heavily dependent upon the quality of the material submitted.

The following explanation of techniques is provided to help clarify markings or notations which may appear on this reproduction.

1. The sign or “target” for pages apparently lacking from the document photographed is “Missing Page(s)”. If it was possible to obtain the missing page(s) or section, they are spliced into the film along with adjacent pages. This may have necessitated cutting through an image and duplicating adjacent pages to assure complete continuity.

2. When an image on the film is obliterated with a round black mark, it is an indication of either blurred copy because of movement during exposure, duplicate copy, or copyrighted materials that should not have been filmed. For blurred pages, a good image of the page can be found in the adjacent frame. If copyrighted materials were deleted, a target note will appear listing the pages in the adjacent frame.

3. When a map, drawing or chart, etc., is part of the material being photographed, a definite method of “sectioning” the material has been followed. It is customary to begin filming at the upper left hand corner of a large sheet and to continue from left to right in equal sections with small overlaps. If necessary, sectioning is continued again—beginning below the first row and continuing on until complete.

4. For illustrations that cannot be satisfactorily reproduced by xerographic means, photographic prints can be purchased at additional cost and inserted into your xerographic copy. These prints are available upon request from the Dissertations Customer Services Department.

5. Some pages in any document may have indistinct print. In all cases the best available copy has been filmed.
BOLT, GREGG ALLEN

PSYCHOMETRIC CHARACTERISTICS OF THE BEHAVIORAL OBSERVATION SCALE

WESTERN MICHIGAN UNIVERSITY M.A. 1984

University
Microfilms
International 300 N. Zeeb Road, Ann Arbor, MI 48106
TABLE OF CONTENTS

ACKNOWLEDGEMENTS ............................................. ii
LIST OF TABLES .................................................... v
LIST OF FIGURES .................................................. v

Chapter

I. INTRODUCTION ..................................................... 1
   Halo Error .......................................................... 11
   Leniency Error .................................................... 16
   Convergent and Discriminant Validity ...................... 19

II. METHOD ............................................................. 21

III. RESULTS ........................................................... 29

IV. DISCUSSION ....................................................... 36

APPENDICES

A. BEHAVIORAL OBSERVATION SCALE FOR PSYCHIATRIC AIDE .......................... 46
B. INSTRUCTIONS FOR BOS ........................................................................ 51
C. ESTIMATES FOR VARIANCE COMPONENTS ........................................... 52
LIST OF TABLES

Table

1. Example of BES .............................................................. 3
2. Example of BOS .............................................................. 4
3. Example of Graphic Rating Scale ................................. 5
4. Percentage of Raters in Each Category for items on BOS . 29
5. Means, Variances, One-Way ANOVA, Levene's Test for Equal Variances for Performance Ratings by Rating Course for Each Unit ........................................... 30
6. Dunn-Bonferroni Comparison Tests for Performance Ratings ................................................... 32
7. Weighted Means for Source Ratings ............................... 34
8. Three-Way Analysis of Variance Summary Table ............ 35
9. Number and Percentage of Supervisors in Each Category: Latham et al. (1979), Data .......................... 38

LIST OF FIGURES

Figure

1. Mean Source Ratings by Psychiatric Unit ......................... 33
CHAPTER I

INTRODUCTION

The assessment of how well people perform on their jobs has been the focus of considerable research and debate over the past 60 years. The academician and the practitioner have generated literally hundreds of articles suggesting new appraisal systems, revising old ones, and bantering over how to control for rater bias and the like. Why such a furor over appraisal systems? The demand for an effective and efficient method for assessing performance arises out of one or more of the following four purposes: (a) appraisals are the basis of promotion and placement decisions; (b) appraisals are frequently used to determine merit allocation; (c) appraisals are the criterion against which selection devices and training programs are validated; and (d) appraisals are one of the primary sources of performance feedback (Kane & Lawler, 1979). If organizations are to successfully utilize performance appraisals as the data base for personnel decisions, then they must be concerned about how well a given performance appraisal accurately reflects actual performance. In light of this concern, the purpose of this study was to explore some of the psychometric properties of a relatively new appraisal system, the Behavioral Observation Scale (BOS), developed by Latham and Wexley (1977). However, before the specifics of the study are described, the concerns for psychometrically sound performance appraisals warrent further comment.

Despite the continuous flow of research and new state of the
Art appraisals, many of the currently used appraisal systems fall short of expectations in terms of discriminant and convergent validity, reliability, and freedom from rater bias (Kane & Lawler, 1979). For those who must rely on performance appraisals for personnel decisions, awareness of the limitations of appraisals compounds an already difficult decision making process. Furthermore, legal requirements for appraisals are enforced by The Equal Employment Opportunity Commission (EEOC), the Office of Federal Contract Compliance Programs, and the courts, who demand validity studies for appraisals contributing to adverse impact. Since the time EEOC wrote "Guidelines for Employment Selection Procedures" (1970), which in effect placed legal requirements on "tests and other selection procedures which are used as a basis for any employment decision" (p. 65543), numerous court cases have been lost by organizations because employers implemented performance appraisals contributing to adverse impact. (For reviews see Cascio & Bernardin, 1981; Schneier, 1978.) As a result of these legal pressures, the demand for psychometrically sound appraisal systems has increased in the working community.

In response to the needs for effective appraisal systems, a number of new models and technical advances in performance appraisals has appeared in the literature over the past 20 years (e.g., Allen & Rosenberg, 1978; Latham & Wexley, 1977; Rosinger, Myers, & Leoy, 1982; Smith & Kendall, 1963). One of the new models is the BOS. Within the literature, the BOS is at times referred to as an extension of the Behavioral Anchored Rating Scale (BARS) or the Behavioral Expectation Scale (BES) developed by Smith and Kendall (1963)
(e.g., Feldman, 1980; Landy & Farr, 1980). The BES differs from the BARS in that BES behavioral statements are written as expectations rather than as neutral behaviors as with BARS (BARS and BES will be used interchangeably). Both BES and BOS are developed using the Critical Incident Technique (Flanagan, 1959). However, the developers of a BES generate behavioral anchors from the incidents, allocate the behavioral anchors to specific dimensions, and use seven of the anchors to form a Thurstone-type rating scale. A BES example is shown in Table 1.

Table 1
Example of BES

<table>
<thead>
<tr>
<th>Motivation: Willingness to work hard</th>
</tr>
</thead>
<tbody>
<tr>
<td>7  Employee could be expected to lend help to other employees when own work is finished</td>
</tr>
<tr>
<td>6</td>
</tr>
<tr>
<td>5  Employee could be expected to organize time to insure completion of tasks</td>
</tr>
<tr>
<td>4</td>
</tr>
<tr>
<td>3  Employee could be expected to need frequent reminders about tasks at hand</td>
</tr>
<tr>
<td>2</td>
</tr>
<tr>
<td>1  Employee could be expected to frequently forget to complete work and report unfinished tasks</td>
</tr>
</tbody>
</table>

On the other hand, the developers of the BOS derive behavioral descriptions from the incidents, allocate the descriptions to specific dimensions, and attach a Likert-type scale to each description within each dimension. An example is shown in Table 2. The specific
procedures for developing a BOS are found in Latham and Wexley (1981). Latham and Wexley (1981) primarily developed the BOS in order to overcome the problems plaguing graphic rating scales, appraisals based on cost-related outcomes, and the BES.

Table 2
Example of BOS

<table>
<thead>
<tr>
<th>Motivation</th>
<th>1. Lends help to other staff when needed</th>
<th>1 2 3 4 5</th>
<th>Almost Never</th>
<th>Almost Always</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>2. Organizes time so all tasks are completed</td>
<td>1 2 3 4 5</td>
<td>Almost Never</td>
<td>Almost Always</td>
</tr>
<tr>
<td></td>
<td>3. Forgets to report unfinished tasks</td>
<td>1 2 3 4 5</td>
<td>Almost Always</td>
<td>Almost Never</td>
</tr>
</tbody>
</table>

Graphic rating scales or trait scales have been increasingly criticized by other researchers and unfavorably looked upon by the courts (Borman & Dunnette, 1975; Holley & Field, 1975; Kleiman & Durman, 1981; Kleiman & Faley, 1978; Latham & Wexley, 1977; Scheier, 1978). An example is shown in Table 3. Briefly, traits tend to be ambiguous and cause confusion and misinterpretation by rater and ratee alike. In addition, at the time of evaluation, unless the evaluator knows specifically what behaviors the traits denote, feedback from trait scales
can frequently be meaningless or misleading, and subsequently have little impact on future performance. Cascio and Bernardin (1981) suggested, "performance dimensions should be behaviorally based. Avoid abstract trait names in graphic rating scales."

Table 3
Example of Graphic Rating Scale

<table>
<thead>
<tr>
<th></th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Unfavorable</td>
<td>2</td>
<td>3</td>
<td>4</td>
<td>5</td>
</tr>
<tr>
<td>Quick Tempered</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Stubborn</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Intelligent</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Firm</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Appreciates Me</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Despite the lack of support for graphic rating scales, it should be noted that little evidence supports psychometric superiority of behaviorally based appraisals over graphic rating scales. In a study designed to assess utility of three rating instruments, including a trait scale and a BARS, Decotiis (1977) reported that the three instruments were approximately equal in terms of their resistance to errors of leniency and central tendency. Landy and Farr (1980) in their review on rating tools concluded, "After more than 30 years of serious research, it seems that little progress has been made in developing an efficient and psychometrically sound alternative to the traditional graphic rating scale" (p. 89). Though one may conclude,
based on common sense or intuition, that behaviorally based appraisal tools are superior to trait scales, data supporting psychometric superiority has yet to be documented.

Appraisals based on cost-related outcomes have some value, but when used in isolation from other performance data, they can often be misleading and omit relevant performance information. Such cost-related measures typically include economic or cost-related outcomes of the organization (e.g., profits, costs, return on investments). Latham and Wexley (1981) cited the following problems associated with cost-related formats: (a) cost-related measures frequently omit relevant factors for which the ratee should be held accountable; (b) cost-related measures are often difficult to obtain for every employee; (c) cost-related measures can for some employees involve factors beyond their control; (d) cost-related measures can fail miserably in providing specific performance feedback necessary for increasing or maintaining productivity; and (e) cost-related measures can foster a "results-at-all-costs mentality" which can run counter to organizational values and goals (pp. 41-44). It seems that if cost-related measures are to be used, they should be carefully scrutinized to reflect only the factors under the control of the ratee and be used as complementary data for behaviorally based data. Latham, Fay, and Sarri (1979) supported the need to behaviorally based data, for without it, "it may be easy to determine whether an employee is or is not meeting a set of objectives, but the answer(s) to the question(s) of how and why can remain elusive" (p. 300).

The BARS has received a considerable amount of attention in the
literature since its development by Smith and Kendall (1963). (For reviews see Landy & Farr, 1980; Schwab, Heneman, & Decotiis, 1975.) It was the intention of Smith and Kendall to develop a behaviorally based rating scale derived from a complete job analysis; hence, the scale would take into account all critical behaviors of a job and be specific enough to avoid the confusion of ambiguous trait names. In addition, the scale could encompass cost-related measures. The BARS was expected to provide psychometrically sound and specific performance feedback. Unfortunately, the BARS has fallen short of original expectations. Studies that have set out to support the psychometric properties of the BARS have reported equivocal results (Bernardin, 1977; Bernardin, Alvares, & Cranny, 1976; Borman & Dunnette, 1975; Campbell, Dunnette, Avery, & Hellervick, 1973; Kingstrom & Bass, 1981; Landy & Farr, 1980; Schwab et al., 1975; Shapira & Shiron, 1980).

Borman and Vallon (1974) reported that after developing a BARS in one setting and using it in another setting, the effectiveness of the appraisal decreased. Subjectivity in categorizing the anchors on a BES is cited as a problem by Latham and Wexley (1981), since nonindependent categories may result in redundancy. Finally, Borman (1979) and Landy and Farr (1980) argued that frequently a rater using BARS has problems discerning the similarity between anchors on the scale and actual performance, which may result in significant rating errors and poor validity. The final question raised by Landy and Farr (1980) concerns whether or not the benefits outweigh the costs of developing a BARS. This appears to be a legitimate concern in light of BARS limitations.
In order to overcome the limitations of BARS, and yet retain the practical and legal advantages of a behaviorally-based rating scale, Latham and Wexley (1977) developed the BOS. In a discussion that cited four disadvantages with BES usage, Latham and Wexley (1981) argued that such limitations do not occur with BOS usage. First, "endorsement of an incident above the neutral point on the BES implies endorsement of all other incidents between the incident checked and the neutral point" (p. 63). The rater using the BOS is allowed to evaluate the ratee on all relevant behaviors within a behavioral dimension; whereas the rater using BES is forced to evaluate the ratee on an entire behavioral dimension with a single endorsement. Problems occur when the rater cannot endorse items between the neutral point and the behavioral item endorsed. This problem does not occur with BOS usage.

Second, "the subjective definition of 'critical' is minimized in the generation of the behavioral items for BOS" (Latham & Wexley, 1981, p. 63). In the process of BES development, only those items judged to be "critical" are retained for anchors on the rating scale, thereby increasing the chances of significantly reducing content validity. Because all behavioral items that are not redundant are retained for the BOS, content validity is not jeopardized in BOS development.

Third, "in using BES, standard or normal behaviors may not be remembered in the same way as unusual or unique behaviors" (Latham & Wexley, 1981, p. 64). In order to overcome this problem the BES user must systematically record performance on normal, routine behavior. The recording procedure could easily become a time consuming task if
the relevant behaviors are unknown. BOS eliminates this problem because it serves as a checklist for both the rater and ratee; irrelevant behaviors are ignored.

Finally, Latham and Wexley suggested that the range of behaviors on a single dimension may be biased by the judges who develop it. Atkin and Colon (1978) conducted research on the Thurstone scale and found that when judges believe one dimension is significantly more important than others, they will describe few acceptable behaviors, many unacceptable behaviors, and almost no neutral behaviors. Problems of this sort are avoided if one uses a BOS. On the BOS a rater is simply required to rate the frequency of behavior observed; all relevant behaviors are found on the scale. Although Latham and Wexley (1981) argued for the superiority of the BOS, most of their arguments were based on logic rather than research data. In assessing the psychometric characteristics of the BOS, Latham and Wexley did not argue that the BOS was superior to the BES, but they did suggest that the scale satisfied EEOC requirements and standards.

The studies that supported Latham and Wexley's contention that the BOS was satisfactory both in terms of reliability and validity for assessing performance, were surprisingly, based on the same data set. Latham and Wexley (1981) wrote: "In previous studies (Latham & Wexley, 1977; Latham, Wexley & Rand, 1975; and Ronan & Latham, 1974) the test-retest and interobserver reliability, as well as the validity of the BOS in indicating employee attendance and productivity, were demonstrated" (p. 63). Although a complete analysis of each study is beyond the scope of the present study, it should be pointed out that
the three studies cited in the above quote have hypotheses supported by the same data set taken from performance of loggers in the Southeastern United States. From this, one might conclude that at best the reliability and validity of the BOS appears promising, but further research is necessary before more conclusive statements can be made.

In addition, Latham and Wexley (1981) contended that the BOS typically satisfies EEOC standards in terms of content validity and interjudge agreement of categorization. These standards will most likely be met if the procedure that Latham and Wexley (1981, 1977) described for BOS development is followed.

With respect to rater bias, Latham and Wexley (1981) suggested that bias was minimized, "because observers do not have to extrapolate from what they have observed to the placement of a checkmark beside an example on the scale that may or may not be appropriate" (p. 63). Empirical support of this final contention has yet to be documented.

The need for further research on the psychometric properties of the BOS is evident. From a psychometric standpoint, the BOS has received only indirect criticism. The criticism focused on behaviorally based performance scales has frequently been directed at the BARS or BES and only indirectly at the BOS (e.g., Landy & Farr, 1980). The contention that the BOS is an extension of the BES does not necessarily allow one to argue that the psychometric properties of the BES are synonymous with those of the BOS because the psychometric properties of the Thurstone scale are not synonymous with those of the
Likert scale. The need, therefore, for independent research on the psychometric characteristics of the BOS is necessary in order that the user of the BOS may be assured of its effectiveness. The research already completed on halo error, leniency error, convergent and discriminant validities will be reviewed.

Halo Error

Halo error has been defined by Holzbach (1978) as a bias in ratings that occurs when a rater evaluates an individual on various items and dimensions without differentiating among them, but instead evaluates the ratee according to a single global or overall judgment. A second usage of halo error was offered by Cooper (1981); he wrote, "Salient features affect the ratings of categories that the rater believes are related to the salient features" (p. 218). Both definitions, though conceptually different, are operationalized similarly. In both cases, the raters are depicted as tarnishing the ratings by evaluating the ratee in light of a global evaluation or of some salient feature(s).

In a review entitled, "Ubiquitous Halo", Cooper (1981) identified two forms of halo error that occur in all rating scales; illusory halo and true halo. Illusory halo is what one generally thinks of as halo error; it is the bias in ratings that most appraisal users wish to avoid. True halo is operationalized as the true correlations that exist between dimensions. Any correlation between two dimensions on an appraisal will consist of some illusory halo and some true halo. The correlation coefficient is the sum of true halo and illusory halo. Both types of halo warrant review.
Five sources of illusory halo have been identified by Cooper (1981): undersampling, engulfing, insufficient concreteness, insufficient rater motivation and knowledge, and cognitive distortions. Undersampling as a source of halo error occurs when the rater has insufficient information on the ratee's behavior; therefore, the rater is forced to rely on a global impression or a few salient features to make each rating decision.

Engulfing traces halo error to the rater's belief that categories covary with global impressions or salient features.

Halo error that is attributable to insufficient concreteness occurs when raters base their ratings on salient features because raters are unable to differentiate item dimensions. An implication of this theory is that if rating dimensions and rating items are highly descriptive as opposed to abstract, halo error will be reduced. Empirical evidence for this has been equivocal. Cooper (1981) found evidence to support the hypothesis that halo error is reduced with highly descriptive and concrete rating items. Finley, Osburn, Dubin, and Jeannert (1977) reached no firm conclusions on the effects of general and specific anchors on their rating scale.

A fourth source of illusory halo attributes biased ratings to one's inability or unwillingness to sensitize oneself to committing halo errors. In an attempt to remove this source of halo error, attempts have been made to train raters to reduce illusory halo. Some methods of training have been more successful than others, however, mixed results have been more prevalent (e.g., Borman, 1979; Fay & Latham, 1982; Thorton & Zorich, 1980; Warmke & Billings, 1979; Zedeck
A final source of illusory halo occurs as the result of cognitive distortions. Cooper (1981) argued that stored observations become distorted over time as one adds and deletes information in the cognitive process. Essentially detail is lost and beliefs about dimension covariance are added. As in the typical organization with an annual review, raters must recall an individual's performance from the past year; actual behaviors cannot be recalled, but impressions resulting from cross-dimension correlations are recalled. If the cross-dimension correlations overstate true covariance, illusory halo results. Cooper wrote, "This fifth source has been unappreciated in the halo-reduction literature" (p. 221).

For those who must rely on rating scales for performance reviews, the presence of true halo in addition to illusory halo make ratings difficult to interpret. The premise that true correlations exist between dimensions has been supported in the literature (Cooper, 1983; Fay & Latham, 1982; Murphy, 1982). Cooper (1981) argued that the abilities to perform a specific job are frequently more homogeneous than heterogeneous. Cooper concluded that although a job may possess various duties, the skills and abilities to perform those duties are often dependent and correlated resulting in true halo on performance appraisals. For the researcher as well as the employer of rating scales, the implications of this would suggest that in order to assess what is true halo and what is illusory halo, actual between dimensions correlations must be computed. Murphy (1982) agreed and wrote, "Unless the researcher has some independent estimate of true
correlations among performance dimensions and of the correlations between performance appraisal items and overall evaluations, it is simply not possible to cite the observed correlations as evidence of rating error" (p. 162).

The evidence that supported the existence of true and illusory halo was convincing and was taken into consideration when assessing halo error in the present study. No attempt was made to suggest a higher or lower magnitude of halo error using the BOS as compared to other appraisal systems. The present study was designed to assess magnitudinal differences in halo errors across raters (self, peer, and supervisor) using the BOS. It was concluded then that what was illusory halo and what was true halo would make no differences in the results of the present study.

As suggested above, the present study was concerned with halo errors as they occurred across raters; a significant amount of research has been conducted which addresses these issues, although none has been found which uses the BOS. (See Landy & Farr, 1980 for review.) Studies which examined differences in halo error across the role of the rater reported equivocal results. Thorntom (1980) reviewed the literature on psychometric properties of self-appraisals and found 12 studies reporting higher incidence of halo for self-ratings vs. peer- and supervisor-ratings; however, he also found 10 studies where self-appraisals manifested less halo than comparison groups. The differences in halo error in peer-ratings vs. supervisor-ratings has also been studied. Klimoski and London (1974) reported a greater degree of halo error in peer-ratings, whereas Holzbach (1978), found similar
degrees of halo in peer-ratings and supervisor-ratings.

Researchers have suggested various hypotheses to explain halo error differences among raters. One hypothesis states that raters who occupy different roles in an organization view any one particular job from different vantage points (Borman, 1974; Holzbach, 1978; Schneier & Beatty, 1978; Zedeck, Imporato, Krausz, Oleno, 1974). Each rater in this case may have different expectations for an individual's performance based on their (rater) own jobs and experiences. Schneier and Beatty (1978) hypothesized that "differences in job duties and proximity, causing differing frequencies and/or duration of observation or ratee performance, could account for divergent ratings given by, for example, superiors and peers" (p. 130).

If the problem of interpretation of appraisal dimensions and items exists for raters occupying different vantage points, then one might expect that if the appraisal items were very specific and behavioral in nature then every rater from each vantage point should interpret the appraisal item in the same way, thus reducing the differences in halo error across raters. After examining the studies that assessed halo error across raters, it was found that many researchers used graphic rating scales or BARS (Borman, 1974; Heneman, 1974; Holzbach, 1978; Klimoski & London, 1974; Lawler, 1967; Lee, Malone, & Greco, 1981; Parker, Taylor, Barret, & Martens, 1959; Schneier & Beatty, 1978; Zammuto et al., 1982). In comparison, Cooper (1983) reported that by using very specific, behavioral rating items, halo error was reduced. If insufficient concreteness promotes halo error, and particularly, causes raters from different roles to commit different degrees
of halo error, then it can be hypothesized that by utilizing a behavioral rating scale like the BOS, differences in halo errors across raters will be virtually nil. The point is that rating scales that consist of items that are clearly delineated, specific, and measurable should not be subject to misinterpretation from any vantage point. Therefore, no differences in halo errors should occur across raters from different roles in the organization. In the present study, it was hypothesized that no differences in degree of halo error would occur across supervisor-, peer-, and self-ratings.

Leniency Error

According to Holzbach (1978), "leniency errors, attributable to specific rating sources, occur when ratings from different rating sources on the same ratee group are significantly different" (p. 579). Latham and Wexley (1981) suggested that negative and positive leniency errors are committed by employers who rate too easy or too hard. Two problems occur with undue leniency errors; one measurement problem and one practical problem.

The measurement problem occurs when leniency errors cause undue restriction of range on the performance ratings which limits the magnitude of the possible relationship between the ratings and other variables of interest (Holzbach, 1978). The practical problem occurs when the ratee interprets the performance ratings. With positive leniency errors, the performer will incorrectly assume adequate performance and continue with poor performance. The appraisal contaminated with negative leniency errors will incorrectly reflect poorer performance than actually occurs. This performer may be deprived of rewards or
promotions that were deserved. In either situation, poor performance may result with undue leniency errors.

With respect to leniency errors as they occur across raters, the evidence suggests that self-appraisals are more lenient than either peer- or supervisor-ratings (Holzbach, 1978; Klimoski & London, 1974; Meyer, 1980; Parker et al., 1959; Schneier, 1978; Thorton, 1980), although one study (Heneman, 1974) reported less leniency error for self-appraisals in comparison to supervisor-appraisals. With respect to supervisor- and peer-appraisals, two studies (Schneier, 1978; Zedeck et al., 1974) reported that supervisor-appraisals demonstrated less leniency errors, and one study (Holzbach, 1978) reported no significant differences.

Very little research has been done to explain why leniency errors occur. Zammuto et al. (1982) reported that organizational differences in leniency error occurred for six items on their performance appraisal though no conclusions were reached as to why this occurred. No doubt many of the rater characteristics discussed earlier affect leniency errors. In addition, leniency errors may be greater simply because an individual wants to make himself/herself appear competent or a fellow employee to appear competent. The consequences of the appraisal then could play an important role; Zedeck and Cascio (1982) supported this assumption. Consistent with the research reviewed, it was hypothesized that self-appraisals would demonstrate more leniency errors than either peer- or supervisor-ratings.

Before the literature is reviewed on convergent and discriminant validity, a discussion on the operational definitions of halo error
and leniency error is warranted, since some have found that different operational definitions yield different values of error (e.g., Saal, Downey, & Lahey, 1980). Saal et al. described four methods of halo error assessment taken from the literature. Briefly, the methods are the following: (a) comparison of mean dimension ratings which examine the intercorrelations among different dimensions, higher intercorrelations suggest greater halo error; (b) the result of factor analysis of the dimension intercorrelation matrix, fewer factors or principle components that emerge are indicative of greater halo error; (c) analysis of the variance or standard deviations of a rater's rating of an individual across each performance dimension, less variance or restricted standard deviations suggests greater incidence of halo error; and (d) rater x ratee x dimension ANOVA, where a significant rater x ratee interaction, especially one that accounts for a sizeable proportion of the total variance, is interpreted as halo error. The operational definition used in the present study to assess incidence of halo error among raters is the third definition, examination of rater variance. Although Saal et al. (1980) criticized all four definitions as poor indicators of absolute halo, for the present purposes of comparing halo error among raters this definition will suffice.

According to Saal et al. (1980) three operational definitions exist for assessing leniency error. The first definition, the most popular one, is to compare the mean dimension ratings with the mid-point of the scale. Mean ratings that significantly exceed the mid-point of the scale reflect leniency; whereas, mean ratings that are below the midpoint of the scale reflect severity. The second definition
suggests a rater x ratee x dimension ANOVA. A rater main effect, especially one that accounts for a large proportion of the total variance, is said to reflect leniency. Finally, Saal et al. suggested a few studies examined the degree of skewness of dimension ratings for evidence of leniency. A significant negative skewness of dimension ratings is said to reflect leniency; whereas, a significant positive skewness is said to reflect severity. A problem with assessing leniency is that without actual performance data, no absolute degree of leniency can be determined. The present study operationally defined leniency as present when supervisor-, peer-, and self-mean ratings differed significantly across behavioral items. Interest was in examining incidence of leniency, not in assessing absolute leniency.

Convergent and Discriminant Validity

Since Campbell and Fiske (1959) first introduced the multitrait-multimethod analysis (MTMM) as a means to assess convergent and discriminant validity, a significant amount of research has been done to assess these validities of appraisal scales. (For reviews, see Holzbach, 1978; Lee, Malone, & Greco, 1981; Kavanaugh, MacKinney, & Wolins, 1971.) Included in that research were investigations reporting on the discriminant and convergent validity of the ratings obtained with BARS. A few studies (Dickenson & Tice, 1973; and Zedeck & Baker, 1972) have reported little of either validities. In contrast, Friedman and Cornelius (1976) found evidence of convergent validity and less halo when participants were active in BARS construction. In addition, Lee, Malone, and Greco (1981) using MTMM, found good convergent and discriminant validity using a summed rating scale. No research was found
assessing discriminant validity and convergent validity of a BOS.

Before the present hypothesis is proposed, perhaps it is worthy to review the meaning and value of both convergent and discriminant validity. Convergent validity has been defined by Holzbach (1978) as, "the extent of agreement between two or more measures of the same trait using different methods" (p. 580). Discriminant validity has been defined by the same as, "the extent of independence between measures of different traits" (p. 580). Although the definitions of convergent and discriminant validity are fairly straightforward the value of assessing them is more obscure. Lawler (1967) wrote, "the primary gain from a research point of view is that this approach [MTMM] allows the researcher to develop a much more sophisticated understanding of his criteria than is possible where it is not employed" (p. 372). Part of this understanding, as explained by Lawler, comes about through determining convergent and discriminant validity. Unfortunately, it seems that, like Lawler, numerous researchers have assumed that more information is better than none, since many of the studies reviewed by the present researcher never mentioned why discriminant and convergent validity were being assessed.

The importance of determining why convergent and discriminant validity are valuable must be accomplished so that the results of any study assessing them can be put into proper perspective. First, Campbell and Fiske (1959) wrote, "validation is typically convergent, a confirmation by independent measurement procedures. Independence of methods is a common denominator among the major types of validity (excepting content validity) insofar as they are to be distinguished.
from reliability" (p. 81). Although convergent validity cannot and does not evaluate validity in the absolute sense, evidence of convergent validity, correlations between the same items or dimensions as rated by various raters being significantly larger than zero, indicates that raters are rating the same construct or behavior, as is the case with a BOS. Therefore, with a BOS, significant convergent validity would suggest that raters (peer, self, and supervisor) are rating the same construct or behavior.

Second, Campbell and Fiske (1959) wrote, "for justification of novel trait measures, for the validation of test interpretation or for the establishment of construct validity, discriminant validation as well as convergent validation is required. Tests can be invalidated by too high correlations with other tests from which they were intended to differ" (p. 81). In the same way, behavioral items on a BOS can be invalidated by too high correlations with other items from which they were intended to differ. If the BOS used in the present study was found to possess some significant degree of discriminant validity, then the present BOS could be said to differentiate among behavioral items. Whether or not differentiation was accurate and meaningful cannot be argued without additional "true" performance data. All that could be argued is that the behavioral items do discriminate performance among ratees and/or within one ratee in an orderly fashion.

It should be summarized again at this time that even if a BOS was found to possess significant degrees of convergent and discriminant validity, direct inferences about the accuracy or criterion-related validity of the items could not be made. It is possible that a BOS
could possess both types of validity for assessing performance, yet be invalid in the sense that it measures "true" performance. True performance data are needed to make true validity inferences. This observation about convergent and discriminant validity was further delineated and supported by Lawler (1967).

Similar to the purposes of the present study, other investigations that have utilized MTMM approach to assess convergent and discriminant validity for combinations of supervisor-, self-, and peer-ratings have generally reported support for convergent validity and little or no support for discriminant validity (Heneman, 1974; Kavanagh et al., 1971; Klimoski & London, 1974; Lawler, 1967; Lee, Malone, & Greco, 1981). Lack of discriminant validity has been primarily attributed to the occurrence of large halo effects. One observation that can be made regarding these studies is that the appraisal scales used were graphic rating scales or BARS. As pointed out earlier, using a summated rating scale with specific items produced good convergent and discriminant validity (Lee et al., 1981).

In the present study, the discriminant validity and convergent validity of the BOS to measure performance were explored using the MTMM. The multitraits were the behavioral items on the BOS. The multimethods were the rater sources, peer, self, and supervisors. It was hypothesized that since the items were very concrete and specific, significant convergent validity and discriminant validity would be found.

In review, the other two hypotheses previously proposed are that (a) no differences in degree of halo error would occur across rater
sources; and (b) self-ratings would demonstrate more leniency error than peer-ratings or supervisor-ratings.
CHAPTER II

METHOD

Subjects

Subjects for BOS development were 29 male psychiatric aides (aides) and 1 female aide randomly selected from a population of 95 aides from a psychiatric facility in Western Michigan.

Subjects who provided data for hypotheses testing included 49 male aides and 2 female aides, six nursing supervisors, and an undetermined number of peers, both aides and nurses (RN's), providing 156 peer ratings. (The number of peers could not be determined, since some peers rated more than one aide and rater's names were kept anonymous.)

BOS Development

The procedure followed for BOS development closely resembled the procedure outlined by Latham and Wexley (1981). From a computer generated list of 95 aides, 30 aides were randomly selected using a table of random numbers. From each of the six psychiatric subunits, a proportion of aides was selected equal to the proportion of aides on the specific unit to the total number of aides. The critical incident technique developed by Flanagan (1959) was utilized to collect ten critical incidents from each aide, five incidents that described effective behavior and five incidents that described ineffective behavior and five incidents that described ineffective behavior for each aide.
aide. (Further clarification of effective and ineffective incidents can be found in Latham and Wexley, 1981.)

A total of 300 incidents was collected. After initial screening for redundancy and ambiguity, a list of 167 incidents was obtained.

Since each of the six units treated different types and ages of patients, aide job duties varied, necessitating the development of five slightly modified BOS's. Relevancy of the 167 incidents to the units was accomplished by the respective nursing supervisor. Each supervisor edited the list for relevant items and appropriate medical jargon. A total of 92 incidents remained.

The next step was to determine overall job categories or behavioral criteria under which the incidents would be grouped (e.g., work habits, staff interactions, communication). Nine of the ten criteria selected were chosen from the existing aide job model. Two aides, one nurse, and the researcher collectively assigned the incidents to the broader behavioral categories. Eighteen items did not seem to fit under the given nine criteria, so Work Habits was selected as a tenth criterion.

Content validity was assessed in two ways. First, it was checked to insure that each accomplishment listed in the job description was represented by an incident. No items were added or deleted. Second, a completed BOS was sent to each supervisor and they were asked to add, delete, and/or edit items to make the appraisal job relevant for the respective unit. Supervisors deleted 4 to 26 items.

From the corrections, five Behavioral Observation Scales were constructed with an item range of 66 to 88. All five BOS's retained
the 10 general behavioral criteria. For items representing effective behavior, (1) almost never and (5) almost always served as anchors on the rating scale. For items representing ineffective behaviors; (1) almost always and (5) almost never served as rating anchors. Therefore, a 5 always represented superior performance. Percents defining the values of 1 to 5 can be found in Appendix B as well as the instructions for completing the BOS.

Procedure

Over a five-month period, evaluations of aide performance were completed. Three sources provided evaluations: peers, self, and supervisors. With respect to the peer evaluations, 25 aides were permitted to choose between nurses and/or aides to complete their evaluations; 27 other aides had peers assigned to rate them by their supervisors. The number of peers reporting data on a single aide varied from two to nine. For the analyses, peer ratings were averaged into a single rating. All ratings were recorded on computer scoring cards.

Analyses

The analyses were computed using the data on 48 behavioral items, those found common to the five BOS's. (See Appendix A for the 48 item BOS).

Prior to running the data analyses for the assessment of halo error, leniency error, and convergent and discriminant validity, histograms were plotted for each of the five psychiatric units (units) across the three rater sources in order to assess for normality of
the data distribution. A normal distribution of data has been shown to be a necessary assumption of the ANOVA model (Hopkins & Glass, 1978). A total of 15 histograms was plotted.

In order to assess convergent and discriminant validity and rater bias for the five units taken collectively, the ANOVA technique described by Kavanagh et al. (1972) and Stanley (1961) was utilized. By using a three-way factorial design, Kavanagh et al. found that very large MTMM matrices could be analyzed with considerably less effort. For example, if the present study assessed convergent and discriminant validity by comparing intercorrelations as was the technique described by Campbell and Fiske (1959), 13,000 intercorrelations would have to be examined and compared! The alternative design allowed for the assessment of convergent validity by testing for significant main effects across aides and for the assessment of discriminant validity by testing for a significant interaction between aides and behavioral items in a 3 x 48 x 52 factorial design, where there were 3 rater sources, 48 behavioral items, and 52 aides.

Included in the ANOVA analysis described by Kavanagh et al. was an assessment procedure for rater bias operationalized as a significant interaction between aides and rater sources. Some researchers have incorrectly referred to rater bias as halo error (e.g., Kavanagh et al., 1971; Holzbach, 1978; Lee et al., 1981). It should be pointed out that rater bias may be due to halo error, but that a significant interaction between aides and rater sources (rater bias) may also be due to leniency error or some other systematic rater bias. Therefore, for comparison purposes with other literature, rater bias was calculated though it was
not considered a measure of halo error.

Differences in halo error among rater sources for each of the five units were compared using Levene's Test for equal variances. Less variance for any one rater source for the 48 behavioral items was indicative of halo error.

Utilizing a one-way ANOVA, leniency error was operationalized as a significant rater mean difference. Leniency error was assessed for each unit. Given a significant F-ratio, the Dunn-Bonferroni (Huitema, 1980) was utilized to determine which rater source was most severe and most lenient.
CHAPTER III

RESULTS

The results obtained from computing the histograms suggested that the data violate the assumption of normality in the ANOVA model. Histograms showed that ratings across the three rater sources clustered toward the upper end of the scale causing a strong negatively skewed distribution. Table 4, a summary of the 15 histograms computed, shows the percentage of raters in each rating category.

Table 4
Percentage of Raters in Each Category for Items on BOS

<table>
<thead>
<tr>
<th>Category</th>
<th>Self</th>
<th>Peer*</th>
<th>Supervisor</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1.2%</td>
<td>0.5%</td>
<td>0.42%</td>
</tr>
<tr>
<td>2</td>
<td>1.7%</td>
<td>0.8%</td>
<td>0.68%</td>
</tr>
<tr>
<td>3</td>
<td>7.1%</td>
<td>4.7%</td>
<td>7.3%</td>
</tr>
<tr>
<td>4</td>
<td>31.2%</td>
<td>22.6%</td>
<td>28.3%</td>
</tr>
<tr>
<td>5</td>
<td>59.2%</td>
<td>71.3%</td>
<td>63.4%</td>
</tr>
</tbody>
</table>

*Peer scores were rounded to the nearest whole number.

Table 5 presents the means and variances for the five units and the results of the one-way ANOVA and Levene's Test for equal variances. Halo error in performance ratings was operationally defined as present when the variance associated with supervisor, peer, and self-ratings
<table>
<thead>
<tr>
<th>Unit</th>
<th>Self</th>
<th>Supervisor</th>
<th>Peer</th>
<th>ANOVA</th>
<th>Levene's</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>M</td>
<td>V</td>
<td>N</td>
<td>M</td>
<td>V</td>
</tr>
<tr>
<td>1</td>
<td>4.373</td>
<td>.669</td>
<td>402</td>
<td>4.284</td>
<td>.945</td>
</tr>
<tr>
<td>4</td>
<td>4.491</td>
<td>.587</td>
<td>501</td>
<td>4.804</td>
<td>.253</td>
</tr>
<tr>
<td>5</td>
<td>4.491</td>
<td>.587</td>
<td>281</td>
<td>4.785</td>
<td>.228</td>
</tr>
</tbody>
</table>

*p .0001
were significantly different or heterogeneous. Significant variance differences by rating sources was found for each of the five units. After further examination of the results, it was discovered that due to the strong negative skewed distributions, a functional relationship existed between source mean ratings and the respective variances. Specifically, a significant negative correlation existed between the rater source means and their respective variance rendering the variance differences uninterpretable as halo error ($r = -0.85, p < 0.01$).

Leniency error in performance ratings was operationally defined as present when mean ratings associated with supervisors, peers, and self-ratings were significantly different. Table 5 presents the results of the one-way ANOVA and Table 6 presents the results of the Dunn-Bonferroni for pair comparisons. The results of the one-way ANOVA demonstrated significant differences among rater sources for each psychiatric unit. Dunn-Bonferroni tests demonstrated that for seven of the eight significant comparisons found between self-ratings and the other two sources, self-ratings were lower or less lenient than both peer and supervisor-ratings.

The comparison tests also demonstrated that of the four significant differences found between peer-ratings and supervisor-ratings, two peer-ratings from two units were higher or more lenient than their respective mean supervisor-ratings and two peer-ratings from two other units were less lenient than their respective mean supervisor-ratings. No difference between peer-ratings and supervisor-ratings were found for the fifth unit.

From the results of the data analyses on leniency error, it
# Table 6

Dunn-Bonferroni Comparison Tests for Performance Ratings

<table>
<thead>
<tr>
<th>Unit</th>
<th>Self vs Supervisor</th>
<th>Self vs Peer</th>
<th>Peer vs Supervisor</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>M</td>
<td>M</td>
<td>F</td>
</tr>
<tr>
<td>3</td>
<td>4.576</td>
<td>4.224</td>
<td>6.069*</td>
</tr>
<tr>
<td>4</td>
<td>4.491</td>
<td>4.805</td>
<td>3.02*</td>
</tr>
</tbody>
</table>

* \( p < .01 \)
appeared as if a relationship or interaction existed between mean source ratings and psychiatric units. Figure 1 demonstrates that, indeed, an interaction existed. From the figure it can be seen that both self-ratings and peer-ratings tend to be more stable than supervisor-ratings across units.

Figure 1. Mean Source Ratings by Psychiatric Unit
In addition to the above analyses, weighted means were calculated for the three rater sources taking the five units collectively. The results are presented in Table 7. The mean self-rating was lower than the peer-rating and the supervisor-rating; whereas, the mean peer-rating was comparable to the mean supervisor-rating.

<table>
<thead>
<tr>
<th>Table 7</th>
</tr>
</thead>
<tbody>
<tr>
<td>Weighted Means for Source Ratings</td>
</tr>
<tr>
<td>Self-rating</td>
</tr>
<tr>
<td>4.398</td>
</tr>
</tbody>
</table>

The results of the ANOVA technique to test for convergent and discriminant validity and rater bias are presented in Table 8. The analysis provided no support for discriminant validity, strong support for convergent validity, and strong support for substantial rater bias. Variance components were also computed. Kavanagh et al. (1971) suggested formulas for estimating variance components so that one might compare the amount of variance due to each source in Table 8. Formulas for computing variance components can be found in Appendix C.
<table>
<thead>
<tr>
<th>Source</th>
<th>df</th>
<th>MS</th>
<th>F</th>
<th>P*</th>
<th>Variance Component</th>
</tr>
</thead>
<tbody>
<tr>
<td>Aide (A)</td>
<td>51</td>
<td>13.964</td>
<td>15.50</td>
<td>.0001</td>
<td>.091</td>
</tr>
<tr>
<td>A x Behavior (B)</td>
<td>2397</td>
<td>1.490</td>
<td>1.65</td>
<td>.16</td>
<td>.196</td>
</tr>
<tr>
<td>A x Source (S)</td>
<td>102</td>
<td>6.585</td>
<td>7.31</td>
<td>.008</td>
<td>.118</td>
</tr>
<tr>
<td>Error (A,B,S)</td>
<td>4794</td>
<td>.901</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

*P = Probability of a Type 1 error.
CHAPTER IV

DISCUSSION

Data Distribution

Preliminary analysis of the data raised some questions as to the appropriateness of the ANOVA models used to assess leniency error and convergent and discriminant validity. Assumptions underlying the ANOVA model include: (a) normality of the data distribution and (b) homogeneity of variances (Hopkins & Glass, 1978). Both of these assumptions were shown to be violated by the data set. Specifically, histograms showed the data set to be negatively skewed; and Levene's Test provided support that the variances associated with rating sources were heterogeneous. Violation of the assumptions has been shown to increase the probability of a Type 1 error.

Despite the violations of assumptions, both ANOVA techniques were used on the data set. In defense of using the ANOVA techniques on the skewed data, research has demonstrated that the ANOVA is robust with respect to violations of the normality assumption given a large sample size (n>30) (Glass, Peckham, & Sanders, 1972; Hopkins & Glass, 1978). The degrees of freedom for the present study, which are directly related to sample size, ranged from 839 to 4794. With respect to heterogeneous variances, the ANOVA model has also been shown to be robust given equal sample sizes of the groups being compared (Glass, Peckham, & Sanders, 1972; Hopkins & Glass, 1978). The sample sizes were approximately
equal in the present data set.

Although it was possible to demonstrate that the ANOVA was robust when the assumptions that underlie the model were violated, other researchers have overcome similar problems by using rating scales that range from seven points (e.g., Friedman & Cornelius, 1976) to 110 (e.g., Lee et al., 1981) and by transforming the data set (e.g., Latham et al., 1979). The possibility of increasing the range of potential responses and transforming the data raised two questions for the present study.

The first question addressed the issue of whether or not a five point scale was adequate for appraising performance. Latham and Wexley (1981) contended that a five point Likert-type scale was adequate for rating scales based on the research of Lissitz and Green (1975) and Jenkins and Taber (1977). A reexamination of this literature proved enlightening.

The research completed by Lissitz and Green (1975) and Jenkins and Taber (1977) involved two Monte Carlo studies which examined the optimal number or rating points for assessing reliability. Both groups of researchers agreed that there was little utility in using more than five rating points given the data are drawn from a normally distributed population. The data used by both groups were generated by a computer where the means of errors, correlations between errors, and correlations between errors and true scores were equal to zero. Jenkins and Taber made this observation:

Although our study was limited to the case in which responses are distributed uniformly across all categories, such distributions are not common in actual research. Future simulations should explore the
generalizability of the current findings for other distributions, especially for the skewed ones often found in applied research. (p. 395)

It seems that Latham and Wexley (1981) failed to realize that performance rating scales often generate skewed distributions thereby making the five point scale undesirable. The data generated by Latham et al. (1979) demonstrated a high degree of skewness evidenced by the need to eliminate 32 of 90 items because the items did not discriminate among performers. Furthermore, even after calculating the total range of scores for each supervisor and dividing by five, the distribution still appeared negatively skewed, as shown in Table 9.

Table 9
Number and Percentage of Supervisors in Each Category: Latham et al. (1979) Data

<table>
<thead>
<tr>
<th>Category</th>
<th>Number</th>
<th>Percentage</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Below Adequate</td>
<td>0</td>
<td>0%</td>
</tr>
<tr>
<td>2. Adequate</td>
<td>0</td>
<td>0%</td>
</tr>
<tr>
<td>3. Full</td>
<td>15</td>
<td>17%</td>
</tr>
<tr>
<td>4. Excellent</td>
<td>59</td>
<td>65%</td>
</tr>
<tr>
<td>5. Superior</td>
<td>16</td>
<td>18%</td>
</tr>
</tbody>
</table>

The question then was raised as to exactly how Ronan and Latham (1974) and Latham, Wexley, and Rand (1975) were able to conduct reliability and validity studies with their obtained skewed distribution. (It should be noted that all validity and reliability studies cited by Latham and Wexley [1981] were generated from the same data set). Ronan and Latham reported interobserver reliability on the raw scores...
was .50 or less for 63 of the 78 behavioral items. Intraobserver reliability was reported as .64 to .84; however, intraobserver reliability was computed by using eight composite scores where the 78 items were subjectively clustered. In the assessment of concurrent validity, Ronan and Latham cited, "Correlations [between items and each criterion] were obtained after normalizing the response to each item" (p. 60). How the authors normalized the data was not reported. Validity coefficients after normalizations ranged from .16 to .31, where 16 of the 17 validity coefficients were significant at the .001 level.

Latham et al. (1975) reported on the intrarater reliability, interrater reliability, and relevance of clustered items taken from a BOS. In order to run the analyses, the 78 behavioral items were grouped into eight criterion scores by taking the algebraic sum of the effective behaviors minus the ineffective behaviors. In other words, the data were transformed in order to increase the reliability and validity coefficients. The point is that both studies, Ronan and Latham (1974) and Latham et al. (1975) demonstrated that without some transformation of the data, the resulting correlation coefficients were lower than desired due to either poor reliability and validity of the BOS for assessing performance or skewness of the data being analyzed. (Skewness has been shown to be detrimental to reliability and validity coefficients by Lemke and Wiersma [1976].) If the low correlations generated from the raw scores were the result of the skewed distribution, then the five point scale would not be the optimal choice for a performance appraisal. The process of transforming the data raised
the second question.

The second question raised concerned whether or not the data generated by the BOS should have been transformed so that the data would be normally distributed. Although three transformations were considered, they were rejected because of the issues they raised.

The first issue dealt with whether or not the original research question, regarding the convergent and discriminant validity of the BOS items, would be answered with transformed data. A transformation considered but rejected was clustering the items together and using the algebraic sum of the items (e.g., Ronan & Latham, 1974). The transformation was rejected on the grounds that the researcher would no longer be assessing convergent and discriminant validity of the items, but of the groups of items.

The second issue dealt with the appropriateness of normalizing the data given no evidence the true population was normally distributed. The two transformations considered were the following: first, using a logarithmic function to transform each data point (recommended by Miller, Note 1); and second, correcting each data point for leniency error (recommended by Brethower, Note 2). The logarithmic function was rejected on the grounds that the resulting statistical analyses from the transformed data would be difficult to interpret (Huitema, Note 3) and, again, no data suggested the true population was normally distributed. Correcting for leniency was rejected because no proof of absolute leniency error existed. It was possible that most aides had excellent performance and that the true population distribution was negatively skewed. Landy and Farr (1980) reached some relevant
conclusions on the research involved in transformations of data generated from rating scales, "in general, this particular area of research raises many more questions than it answers" (p. 92). Future research that would focus on data transformations should take into account, not only how the data are clustered, but also how to interpret the clustered data.

Halo Effects

A secondary problem with the skewed distribution was that it introduced a negative correlation between rater mean ratings and their respective variances. The correlation was such that as mean ratings increased, variances decreased. The correlation had the effect of rendering Levene's Test for equal variances uninterpretable, which in turn made the assessment of halo error impossible. Although Levene's Test were significant for all units across raters, it did not make sense to interpret any one group as demonstrating greater halo error than another group because the variances were negatively correlated with their respective means. Implications for future research would be that one should not use the comparison-of-variance technique for the assessment of halo error if skewed distributions are expected or obtained.

Leniency Effect

Contrary to most previous research, self-ratings tended to be more severe than either peer-ratings or supervisor-ratings. In addition to this study, Heneman (1974) was one of the few researchers...
to find self-ratings less lenient than other rating sources. Heneman hypothesized that self-ratings tended to be more severe because no consequences were made contingent on low ratings. Self-raters in the Heneman study knew that Heneman was collecting data for research purposes only and that no consequences were attached to their ratings. Few if any consequences were attached to low ratings for the BOS used in the present study. (The researcher is not aware of any formal consequences.)

A second observation made from the analysis completed on leniency was that an interaction existed between raters and psychiatric units. The most pronounced interaction involved the supervisor-ratings. Both peer- and self-ratings tended to be stable across units. However, the supervisor tended to fluctuate more between units suggesting more inconsistency in their ratings.

Several explanations could possibly account for the fluctuation in supervisor-ratings and the lack of fluctuation in peer and self-ratings. First, because only one supervisor represented each unit and a number of individuals represented each self-means rating and peer mean rating, any bias in supervisor-ratings would not have been washed out in the averaging process as is possible with the other groups. The effect would be that the supervisor-ratings would appear to be more inconsistent than the other groups. Second, according to Borman (1974), different raters tend to view a single job differently and, therefore, tend to rate inconsistently. It was hypothesized that this effect would not occur since the BOS consisted of behavioral items. However, it is possible that the supervisors still interpreted the
behavioral items differently; unfortunately, there was no systematic way to assess if this was the case. Third, supervisors may have ignored or have been unable to judge the percentage of times an aide engaged in any one behavior. The effect would be that each supervisor used his/her own interpretation of the number of anchors attached to each item on the BOS. The likelihood of supervisors using their own rating interpretations is fairly good, since being able to estimate the percent difference associated with a 4 or 5 appears to be a difficult task given many opportunities to behave in a given manner as dictated by the BOS. Unfortunately, it is not possible to assess which effect or combination of effects caused the rating fluctuations. Future research might focus on each cause in a more controlled setting.

MTMM Interpretation

The results obtained from the three-way ANOVA technique to assess convergent and discriminant validity and rater bias should, according to Kavanagh et al. (1971), be interpreted in the following manner. Differentiation existed among aides attributable to the BOS used, that is, person variance or convergent validity. However, the equally large aide x source effect indicated a substantial method bias confounding the first result. In other words, the ratings for various aides were not consistent across supervisors, or the ratings an aide received were dependent upon which supervisor the aide had as a rater, which would tend to decrease the aide main effect. The lack of aide, x behavior interaction indicated no ordering of aides differently on different behaviors, i.e., no discriminant validity. The lack of
discriminant validity has generally been attributed to substantial rater bias in performance ratings (Heneman, 1974; Holzbach, 1978; Lawler, 1967; Lee et al., 1982).

The important conclusion to be drawn from these results is that the ratings obtained from using an instrument like the BOS, which has specific behavioral items, are not immune to the problems that plague BARS, BES, and graphic rating scales. Such problems include, but are not limited to, halo error and leniency error which adversely effect discriminant validity and generate skewed distributions making analyses difficult to complete and interpret.

These observations regarding the psychometric characteristics of a BOS raised a final question: Should performance rating scales even be used? Landy and Farr's (1980) conclusion is reiterated: "After more than 30 years of serious research, it seems that little progress has been made in developing an efficient and psychometrically sound alternative to the traditional graphic rating scale" (p. 89). Even the BOS, which appears to be the very best attempt to overcome the shortcomings of ambiguous performance rating scales, was found to be no better than the graphic rating scale. Murphy, Martin, and Garcia (1982) concluded, based on their research with the BOS, "the BOS as typically used, measure traitlike judgements rather than behavioral observation" (p. 562).

After completing the literature review on performance appraisals, and the data analysis, it became increasingly evident that true halo error could not be assessed without true performance measures, that leniency error could not be assessed without true performance
measures, that all criterion-related validity studies could not be completed without true performance measures, and the decision to normalize the data could not be made without true performance data that would support a true normalized distribution. Since true performance measures are necessary to compute many of the psychometric properties of rating scales, it seems logical to use them in evaluating performance rather than subjective rating scales. Perhaps if we had spent 30 years of serious research on techniques of obtaining true performance measures rather than attempting to improve rating scales, performance appraisals would be superior to appraisals now available.

Future research should investigate measures of productivity that include outcome measures and process measures of productivity that are not cost related. Latham and Wexley (1981) argued that cost-related measures often omit important information, are difficult to obtain, include factors beyond the performer's control, lead to a results-at-all-costs mentality, and provide inadequate feedback necessary to correct performance. Future investigations should be directed at overcoming these barriers and engineering new methods to measure productivity. Overcoming the hurdles to obtaining true performance measures may be a more productive avenue for research than attempting to overcome the hurdles of subjective rating scales.
APPENDIX A

BEHAVIORAL OBSERVATION SCALE
FOR PSYCHIATRIC AIDE

Interpersonal Relationships with Staff

1. Offers assistance to nurses
   almost never 1 2 3 4 5 almost always
2. Attends staff meetings when possible
   almost never 1 2 3 4 5 almost always

Interpersonal Relationships with Patients

3. Takes initiative to introduce himself to new patients
   almost never 1 2 3 4 5 almost always
4. Participates in daily activity with patients
   almost never 1 2 3 4 5 almost always
5. Spends too much time in the office avoiding interaction with patients
   almost always 1 2 3 4 5 almost never
6. Discusses patient issues confidentially when possible
   almost never 1 2 3 4 5 almost always
7. Is willing to discuss patients complaints
   almost never 1 2 3 4 5 almost always
8. Praises patients for accomplishments
   almost never 1 2 3 4 5 almost always

Communication Process

9. Charts and records 1:1's
   almost never 1 2 3 4 5 almost always
10. Completes his share of the daily charting
    almost never 1 2 3 4 5 almost always
11. Charts in a concise manner
   almost never 1 2 3 4 5 almost always

12. Charts legibly and with good grammar and spelling
   almost never 1 2 3 4 5 almost always

13. Charts with appropriate charting symbols
   almost never 1 2 3 4 5 almost always

14. Makes assessments relevant to treatment plan when charting
   almost never 1 2 3 4 5 almost always

15. Conveys accurate information to team
   almost never 1 2 3 4 5 almost always

16. Listens to patients and staff with attention
   almost never 1 2 3 4 5 almost always

17. Informs at least one other staff member before leaving the unit
   almost never 1 2 3 4 5 almost always

18. Calls ahead if he expects to be late for work
   almost never 1 2 3 4 5 almost always

Spiritual Values

19. Shares spiritual needs and spiritual issues appropriately
   almost never 1 2 3 4 5 almost always

Leadership Abilities/Role Modeling

20. Carries out all delegated tasks
   almost never 1 2 3 4 5 almost always

21. Makes individual decisions when necessary
   almost never 1 2 3 4 5 almost always

22. Is able to continue to function appropriately in stressful situations
   almost never 1 2 3 4 5 almost always

23. Demonstrates good empathic skills
   almost never 1 2 3 4 5 almost always
24. Knows unit rules
   almost never 1 2 3 4 5 almost always

25. Fails to confront when unit rule is broken
   almost always 1 2 3 4 5 almost never

26. Uses humor appropriately
   almost never 1 2 3 4 5 almost always

27. Complains about work, patients, and/or staff
   almost always 1 2 3 4 5 almost never

Educational

28. Attends assigned inservices
   almost never 1 2 3 4 5 almost always

29. Takes opportunity to involve himself in nonmandatory inservices
   almost never 1 2 3 4 5 almost always

Teaching

30. Is able to teach conflict resolution and does so when necessary
   almost never 1 2 3 4 5 almost always

31. Assists in orientating new nursing personnel
   almost never 1 2 3 4 5 almost always

Implementation of Treatment Programs

32. Writes treatment programs with specific objectives
   almost never 1 2 3 4 5 almost always

33. Follows treatment plan issues in 1:1's
   almost never 1 2 3 4 5 almost always

34. Does his share of the close observations
   almost never 1 2 3 4 5 almost always

35. Completes close observations on time
   almost never 1 2 3 4 5 almost always
36. Attends team meetings when possible  
   almost never 1 2 3 4 5 almost always

37. Provides suggestions for treatment plan in team meetings  
   almost never 1 2 3 4 5 almost always

38. Follows team decisions  
   almost never 1 2 3 4 5 almost always

<table>
<thead>
<tr>
<th>Ethical and Legal Issues</th>
</tr>
</thead>
</table>

39. Knows location of safety equipment  
   almost never 1 2 3 4 5 almost always

40. Knows fire and tornado procedures  
   almost never 1 2 3 4 5 almost always

41. Discusses confidential information with patients, families, friends, or relatives of patient  
   almost never 1 2 3 4 5 almost always

<table>
<thead>
<tr>
<th>Public Relations</th>
</tr>
</thead>
</table>

42. Realizes that their communication about their work influences the community's concept of psychiatric care  
   almost never 1 2 3 4 5 almost always

43. Takes an active role as host or hostess to new personnel, patients, visitors, students, and interns  
   almost never 1 2 3 4 5 almost always

44. Can direct others to appropriate persons when additional information is requested  
   almost never 1 2 3 4 5 almost always

<table>
<thead>
<tr>
<th>Work Habits</th>
</tr>
</thead>
</table>

45. Generally displays a positive attitude  
   almost never 1 2 3 4 5 almost always

46. Evidence in mood and attitude that personal problems are not interfering with job performance  
   almost never 1 2 3 4 5 almost always

---

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.
47. Demonstrates a willingness to accomplish various tasks and special projects
   almost never  1  2  3  4  5 almost always

48. Comes to work on time
   almost never  1  2  3  4  5 almost always
APPENDIX B

INSTRUCTIONS FOR BOS

Rate the individual whom you are evaluating on the 5 point scale by darkening the corresponding number on the computer sheet. Rate the employee as best you can in the following manner:

Employees receive a 1 if you suspect they engage in this behavior 0-50 percent of the time, 2 for 50-65 percent of the time, 3 for 65-80 percent of the time, 4 for 80-90 percent of the time, and 5 for 90-100 percent of the time. If you are unable to make a fair rating, leave it blank.

NOTE: the words almost always and almost never are reversed for those items that are worded as inappropriate behavior. Hence, the individual will always be rated a 5 when he is exhibiting excellent behavior.

REMINDE RS: Do not write on this booklet.
Always use a #2 pencil; do not make any stray marks on the computer sheet.
When you erase, erase completely.
Be sure you begin with number 1 on the computer sheet.
Do not fill in any names.
### APPENDIX C

**ESTIMATES FOR VARIANCE COMPONENTS**

<table>
<thead>
<tr>
<th>Source</th>
<th>Variance Formula</th>
</tr>
</thead>
<tbody>
<tr>
<td>Aide (a)</td>
<td>$\frac{MS_A - MS_{A \times B \times S}}{nm}$</td>
</tr>
<tr>
<td>A x Behavior (B)</td>
<td>$\frac{MS_{A \times B} - MS_{A \times B \times S}}{m}$</td>
</tr>
<tr>
<td>A x Source (S)</td>
<td>$\frac{MS_{A \times S} - MS_{A \times B \times S}}{n}$</td>
</tr>
<tr>
<td>Error</td>
<td>$MS_{A \times B \times S}$</td>
</tr>
</tbody>
</table>

Note: $n$ = number of behaviors, $m$ = number of sources

---

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.
REFERENCE NOTES

BIBLIOGRAPHY


