3-1977

A Psychometric Comparison Between Inhouse Versus Externally Developed Retranslation Scales

David Peak

Follow this and additional works at: http://digitalcommons.wku.edu/theses

Part of the Higher Education Administration Commons, and the School Psychology Commons

Recommended Citation

http://digitalcommons.wku.edu/theses/1818

This Thesis is brought to you for free and open access by TopSCHOLAR®. It has been accepted for inclusion in Masters Theses & Specialist Projects by an authorized administrator of TopSCHOLAR®. For more information, please contact topscholar@wku.edu.
A PSYCHOMETRIC COMPARISON BETWEEN INHOUSE VERSUS EXTERNALLY DEVELOPED RETRANSLATION SCALES

A Thesis
Presented to
the Faculty of the Department of Psychology
Western Kentucky University
Bowling Green, Kentucky

In Partial Fulfillment of the Requirement for the Degree Master of Arts

by
David Edward Peak
March 1977
A PSYCHOMETRIC COMPARISON BETWEEN INHOUSE VERSUS EXTERNALLY DEVELOPED RETRANSLATION SCALES

Recommended

[Signature]
Director of Thesis

[Signature]
Dean of the Graduate College

Approved 4-17-77
Acknowledgments

The author is indebted to friends and associates, too numerous to mention, who have given him the moral support and encouragement that he has needed to develop this thesis. Special mention, however, belongs to the following people:

To Ray Mendel, my major advisor, whose guidance and understanding have been crucial to the development of this thesis and will be forever remembered.

To John O'Connor and Sam McFarland, members of my thesis committee, whose criticisms, comments, and suggestions have been helpful.

To Robert Stone, my typist, whose dedication to work has been greatly appreciated.

To my parents, Selby and Arla, whose patience and fortitude have given me the direction in life that I follow now.
# TABLE OF CONTENTS

Chapter 1. Literature Review ........................................ 1
   A Rationale for Judgmental Evaluation ......................... 1
   The Behavior Expectation Scale ............................... 5
   Empirical Comparisons of the BES Format
     to Other Evaluation Formats ............................. 7

Chapter 2. Statement of the Problem ............................. 15

Chapter 3. Method ............................................... 18
   Sample .................................................................. 18
   Scale Development Phase ................................... 18
   Scale Evaluation Phase ....................................... 18
   Procedure ....................................................... 18
     Stage 1 ...................................................... 20
     Stage 2 ...................................................... 21
     Stage 3 ...................................................... 21
     Stage 4 ...................................................... 21
     Final Development ........................................ 22
     Developmental Characteristics of
       Berkeley and Local Forms ............................ 24
     Administration ............................................. 24
   Data Analysis .................................................. 26
     Leniency Effects ............................................. 26
     Discriminability ............................................. 26
     Variability ...................................................... 27
     Interrater Reliability .................................... 28
Chapter 4. Results ................................................. 29
  Effects of Leniency ............................................ 29
  Effects of Discriminability .................................... 31
  Effects of Variability ......................................... 31
  Effects of Interrater Reliability ............................... 32

Chapter 5. Discussion ............................................. 34
  Evaluation of Scales in an Absolute Sense ................. 34
  Evaluation of Scales in a Relative Sense .................. 37
  Comparison between the Original Berkeley Form and the Borrowed Berkeley Form ............ 39
  Comparison between the Local Form and the Original Berkeley Form ......................... 40
  Implications of Results ....................................... 41

Appendix ......................................................... 45
  Berkeley Form .................................................. 47
  Local Form ..................................................... 56
  Bibliography .................................................... 66
Abstract

A PSYCHOMETRIC COMPARISON BETWEEN INHOUSE VERSUS EXTERNALLY DEVELOPED RETRANSLATION SCALES

David Edward Peak March 1977 70 pages

Directed by: Ray Mendel, John O'Connor, and Sam McFarland

Department of Psychology Western Kentucky University

Using a Behavioral Expectation Scale (BES) format borrowed from the psychology department at the University of California—Berkeley and a BES form developed in and for a Southeastern university psychology department, students' evaluations of their professors' teaching performance were examined for interrater reliability, leniency error, variability, and discriminability. Results indicate that neither form was psychometrically sound at the Southeastern university. Problems in obtaining sound ratings for the BES form were discussed.
Chapter 1. Literature Review

A Rationale for Judgmental Evaluation

Responsible persons within educational institutions must confront making critical decisions concerning the quality of instruction given to students. Yet, attaining an understanding of what constitutes effective teaching is a very arduous task. The complexity becomes apparent when one considers that each instructor has a unique teaching style and a unique degree of difficulty in course preparation. In addition, students have differing abilities to grasp material. Without a reference tailored for making decisions concerning teaching effectiveness, identifying who within the university is successful would be arbitrary (Oberg, 1972). If an evaluation successfully focuses upon the goals of teaching effectiveness considered relevant to the educational institution, i.e., has construct validity, then the appraisal instrument can serve as the aforementioned reference (Cummings & Schwabb, 1973).

Selection of a proper evaluation format is a very complex problem also. Most difficult is locating a form which defines the teaching task comprehensively. As noted by Ghiselli (1956), measuring job proficiency can rarely be accomplished using only one dimension. Attempts to measure
teaching proficiency on a single dimension of job success usually result in deficient descriptions of an instructor's total contribution to the class. This deficiency only serves to perpetuate the inadequate rating of the instructors.

In order to make a proper selection of an instrument in light of the problems previously mentioned, several criteria were proposed. Evaluation instruments were analyzed in relation to these criteria (referred to above and discussed in detail below) in order to select an evaluation form most appropriate for analyzing the instructor's performance. The criteria for selecting a psychometric instrument will now be discussed.

First, an appraisal instrument should focus upon human judgment as well as objective indices of performance. Objective performance measures, such as the number of students in class, only partially reflect an instructor's contribution in class. Yet, objective indices are relied upon because few administrators have the time and energy to evaluate each instructor thoroughly. Relying on students who readily observe the instructor in class may be a logical alternative to this problem. Students can provide meaningful feedback about areas of performance readily observable but not necessarily tied to objective measures (Cummings & Schwabb, 1973).

Second, an appraisal instrument should be oriented to observations that are stable and relevant to performance (Campbell, Dunnette, Lawler, & Weik, 1970). If effective teaching performance differs for each rater, one may argue
that evaluative judgments are unreliable and are based on global impressions. Baroff (1954) investigated the degree to which raters could judge the effectiveness of officer candidates in military school. Specifically, a primary objective entailed determining if raters could evaluate officer potential with reliability. Although there were individual differences among raters in their capacities to judge officer potential, the use of several raters cancelled the effects of individual biases, resulting in an internally consistent set of ratings. Psychometric utility of an appraisal system is thus contingent upon a consensus of several raters. Basing an evaluation upon one observation is a questionable practice that should be avoided (Baroff, 1954).

Third, an appraisal instrument should be structured and quantitative rather than composed of qualitative essay statements. Appraisals of teaching which hinge upon open-ended questionnaires about classroom performance focus upon impressions each student holds about an instructor. Evaluating these questionnaires requires inferences to be drawn from essay responses. A structured appraisal system provides a standard set of readily observable items which all raters can use. A common set of standards for evaluation allows better stability in ratings (Jenkins, Nadler, Lawler, & Camman, 1975). Moreover, quantitative evaluations lend themselves to statistical comparisons with alternative rating scales.

Fourth, an appraisal instrument should have raters identify relevant areas of performance by focusing upon specific
behaviors. Few appraisal devices reflect areas of performance systematically. For example, the Student Instruction Report (SIR) measures an instructor's performance across six independent factors (Costin, Greenough, & Menges, 1971). Even though these dimensions of performance were identified through factor analysis and have fine psychometric properties, the raters have not been involved in developing the dimensions and may well not understand them. By taking the time to identify dimensions of performance, the raters direct their efforts toward understanding the multidimensional nature of the teaching task and toward identifying areas of performance relevant to students.

Finally, an appraisal instrument should be an aid to raters in evaluating teaching effectiveness. Forced choice measures frequently violate this requirement. A forced choice scale is a structured instrument in which certain items detect social desirability bias. Such statements serve to trick rather than to help the rater in evaluating performance (Smith & Kendall, 1963). Gordon and Stapleton (1956) found trends which suggest that students give biased responses on a forced choice scale rather than use it critically. Difficulties in understanding form terminology and variable motivation on the part of the raters have been given as explanations for unfavorable findings. A more relevant criticism of this technique may be that the forced choice form confuses the rater.
Finally, an appraisal instrument should provide feedback about performance to the instructor. Blood (1974) found that spinoffs from Behavior Expectation Scale (BES) procedures are well suited for this purpose. Rather than stressing a set of skills to be attained, a BES provides a set of potentially observable behaviors which reflect successful teaching skills. The specificity of behaviors is a great strength for the BES. Not only are the behaviors clearly identified, but students can also evaluate an instructor's performance on each incident with potential performance guidelines. Desirability ratings serve to aid an instructor in discriminating between behaviors which impede and behaviors which reinforce effective teaching performance. Review of these ratings permits the instructor to recognize, within his own repertoire, behaviors resulting in unfavorable performance and behaviors leading to good performance. Therefore, the instructor can use BES devices to further develop his teaching skills.

The Behavior Expectation Scale

The Behavior Expectation Scale (BES) satisfies the criteria for an adequate appraisal instrument as discussed above. The BES is a structured rating form which utilizes a series of graphic scales containing discrete response modes. Formulated by untrained raters participating in a series of structured tasks, the BES utilizes a rigorous developmental procedure. First, developers identify areas of performance which are considered to be relevant to teaching effec-
tiveness. Second, developers generate potentially observable behaviors representing the performance dimensions. In order to ensure that the behaviors represent these dimensions, a third step is included which involves retranslating the behaviors to the performance dimensions. Retranslation is a process where student raters match potentially observable behaviors to the performance dimensions. Only those behavior-performance matches agreed upon by a majority of students are retained. Finally, the desirability of the behaviors is rated quantitatively. A mean rating among the developers defines the point at which each behavior will anchor the scale of the associated performance dimension. In order to ensure that the behavior represents a consensus among students, only behaviors with desirability ratings below a 1.5 standard deviation are retained. In summary, the developmental process of a BES allows students to provide input in the construction phase of the evaluation instrument. Not only is the task ego-involving, but the developmental procedure also circumvents the problem of using areas of performance which do not reflect the patterns of behavior that are perceived to be manifested by the instructors.

BES formats have been used in several situations. As indicated by Bernardin, LaShells, Smith, and Alvares (1976), the BES format characteristically has utility in measuring overt areas of performance across a wide variety of occupations. BES measures have been developed for nurses in hospital settings (Smith & Kendall, 1963), for airline customer
service agents (Campion, Greener, & Wernli, 1973), as measures of motivation (Landy & Guion, 1970), and in the evaluation of teaching performance (Harari & Zedeck, 1973; Burnaska & Hollman, 1974; Bernardin, 1975; and Zedeck, Jacobs, & Kafry, 1976).

Empirical Comparisons of the BES Format to Other Evaluation Formats

Psychometric soundness of a rating device is of paramount importance if performance is to be adequately evaluated. To obtain sound results, the actual behaviors emitted in class by instructors must be recorded by the raters. Accomplishing such a task requires discerning specific behaviors which discriminate among favorable, moderate, and unfavorable performance. Moreover, these accounts of the instructors' performances must be agreed upon by the judges. In short, sound results follow from consistent accounts of teaching performance.

Such extraneous factors as bias and error detract from the soundness of the ratings. Bias can be reflected in lenient descriptions concerning the instructor's performance in class and in composite descriptions rather than ratings which focus on specific behavioral entities. Error can appear in lack of interrater reliability.

Since human behavior is a complex phenomenon, total accuracy in the description of an instructor's performance is rarely obtained. Yet, any evidence of bias or of error
within the ratings detracts from the value of the instrument as an evaluation measure. Since absolute soundness of a measure is rare, the utility of psychometric evaluations is contingent upon how closely they approximate this end goal. By comparing a variety of evaluation forms, one can select the rating device least subject to bias and error. The resultant rating form may lack absolute soundness; if so, a relatively better instrument can be utilized for subsequent evaluations.

Even though the BES has been applied in several settings, only recently has the BES format been systematically compared with other rating formats. Developing comparisons between formats is essential in order to test the soundness of the BES format relative to other instruments. Recent investigations comparing judgmental evaluation formats have produced ambiguous results (Campbell, Dunnette, Arvey, & Hallevik, 1973; Borman & Vallon, 1974; Burnaska & Hollman, 1974; Borman & Dunnette, 1975; Keveaney & MacGann, 1975; and Bernardin, LaShells, Smith, & Alvares, 1976).

One of these formats, the Likert, is criticized for producing only a global index of teaching performance. In contrast, Smith and Kendall (1963), in the development of the BES, asserted that their instrument measures precise areas of performance. Campbell, Dunnette, Arvey, and Hallevik (1973) attempted to determine the relative utility of a BES rating over the Likert. Both formats were used to evaluate the performance of department managers. To develop the BES, depart-
ment managers generated areas of performance as well as behavioral examples of each. They then retranslated the behaviors to performance dimensions, rating the desirability of the behaviors and thereby setting behavioral anchors. To develop the Likert scale, definitions corresponding to each area of performance on the BES were assigned numbers on a discrete scale. Comparing the two formats, the BES was found to be more precise than the Likert in measuring specific areas of performance (more complex factor structure), but the Likert was found to have more consensus than the BES among raters concerning the performance of department managers. This finding suggests that the BES had lower reliability than the Likert.

This study by Campbell, Dunnette, Arvey, and Hallevik suggests that the BES has a structure which facilitates a rater's evaluation of performance on independent dimensions. Clearly, a most important question is the determination of those properties unique to the BES format which favor the BES over the Likert in measuring precise areas of performance.

Borman and Vallon (1974) compared a Likert and a BES format neither of which was specifically developed for the criterion situation. In their attempt to determine which instrument would have relatively greater utility in the evaluation of hospital administrators, Borman and Vallon found that, when neither format was developed for the situation, the raters evaluated personnel on a global impression rather than on independent dimensions of performance. Moreover,
both formats were found to be unreliable. A rather intriguing finding showed that the Likert ratings were less lenient than those of the BES when neither form was developed for the setting.

One may conclude that intensive onsite development by raters of the evaluation instrument is critical when assessing performance. Using actual raters to develop an evaluation instrument results not only in a set of criteria relevant for the criterion situation, but also in raters who are more knowledgeable about the importance of the evaluation. Keveaney and MacGann (1975) attempted to test if rater participation in the development of scales makes a critical difference in the soundness of an evaluation. Specifically, Keveaney and MacGann were concerned with evaluating the performance of university professors using forms having similar structures. In fact, the only structural difference between the two forms was that one had anchors defined by the raters and the other had anchors defined by the psychometrist. The BES with anchors defined by the raters was found to evaluate instructors on independent dimensions of performance, i.e., to possess better discriminant validity, while the BES with anchors defined by the psychometrist yielded evaluations which only globally measured the performance on the instructor.

Even though participation in the development of rating scales may be a critical requirement for a sound evaluation, participative development of rating scales requires a consid-
erable expenditure of effort. Considering the energy expended to develop the form, one would expect it to be superior to other formats. Borman and Dunnette (1975) found that this is not necessarily the case. A trait format, a part of the Naval Officer Fitness report, was compared with a BES format in evaluating officer potential of naval candidates. Being intensively developed for this evaluation, the BES should reflect officer potential more than the trait form traditionally used at the academy. Although the BES was found to have relatively better psychometric utility than the trait format, differences between the ratings were found to be relatively small.

On the other hand, rater participation in the development of scales may produce some ancillary beneficial results. BES forms can provide information to develop training programs, specify goals within organizational policy, and identify communication problems from one unit of organization to the next (Blood, 1974; and Zedeck, Jacobs, & Kafry, 1976). When BES formats are being constructed, numerous examples of effective and ineffective teaching behavior are being generated by developers. Zedeck, Jacobs, and Kafry (1976) have found that several evaluation forms can be derived from these examples. They conducted a study in which two forms, having parallel properties, were compared for evaluative purposes. Specifically, evaluation forms having the same areas of performance but different behavioral anchors were developed to measure teaching performance. Both forms were found to have
equivalent psychometric properties. Also, both forms tended to have lenient ratings in the favorable direction for instructors.

When making comparisons between evaluations, two forms may be equivalent in soundness yet still be inappropriate for rating purposes. Comparisons between forms are made in order to select a rating device which possesses relatively greater soundness. Burnaska and Hollman (1974) assert that judgmental evaluations are unsound due to small relative differences found when comparing ratings. Teaching performance of faculty was evaluated on both absolute and relative psychometric soundness across three formats: (a) a Behavior Expectation Scale with anchors, (b) a Likert format containing areas of performance similar to the BES, and (c) a trait format with categories predefined by the psychometrist. Across all three formats, lenient descriptions and composite ratings of instructors were found to be prevalent. Even though relative differences between formats favored the BES, the psychometric distinctiveness between the forms was slight. As noted by Bernardin (1975), neither the Likert nor the trait format was systematically developed specifically for the evaluation setting. Intensive development should have reduced the lenient descriptions and composite ratings across all the evaluations. Although relative soundness may have favored the BES, bias and error were present in both BES and Likert scales.

From the studies previously mentioned, the results have not been impressive for the BES format. Reliability was
found to be questionable in some studies (Campbell, Dunnette, Arvey, & Hallevik, 1973; and Borman & Vallon, 1974). Relative differences in soundness between the BES and other evaluation formats were found to be small (Burnaska & Hollman, 1974; and Borman & Dunnette, 1975). Moreover, the literature contains both arguments for development and use of judgmental evaluations (Campbell, Dunnette, Arvey, & Hallevik, 1973; Keveaney & MacGann, 1975; Bernardin, 1975; and Zedeck, Jacobs, & Kafry, 1976) and arguments for removal of judgmental evaluations due to substantial bias and error (Borman & Vallon, 1974; Burnaska & Hollman, 1974; and Borman & Dunnette, 1975).

Bernardin (1975) and Bernardin, LaShells, Smith, and Alvares (1976) have attempted to reconcile these conflicting recommendations. Developing several evaluation formats measuring teaching performance, Bernardin's research focused upon isolating critical components in format development and utilization in order to improve the relative psychometric utility of the BES format over other structured evaluations. Critical component comparisons could provide a basis for developing a BES form with greater absolute soundness.

In the first part of the study (Bernardin, 1975), a BES format was compared with two Likert scales. As in the study by Campbell, Dunnette, Arvey, & Hallevik (1973), the first Likert format was developed by selecting all the definitions from the BES format across all performance dimensions. Each definition was changed to a statement and rated on a Likert scale. Going beyond this study, the second Likert format was
intensively developed from the first. Only those definitions which reliably related to performance dimensions were kept for the second Likert format. Intensively developed evaluation formats were found to result in fewer lenient ratings than the evaluation formats not intensively developed. Also, the intensively developed Likert format was more reliable and less subject to leniency error than the BES format.

Bernardin, LaShells, Smith, and Alvares (1976) postulated that the BES format sometimes yields poorer psychometric utility than other structured evaluations because of inadequate scale development and format preparation. Although the literature describes a common theoretical method for developing a Behavior Expectation Scale, critical differences in the actual development and utilization of the BES format affect the scale's psychometric utility. Specifically, Bernardin found that BES formats are most effective psychometrically when there are several anchors widely dispersed along each scale for all performance dimensions. Developing several variations of the BES format, Bernardin found three procedures which produce a wide dispersion of anchors across a scale: (a) using groups of raters with specialized tasks in the developmental procedure of the BES scale, (b) using clarifying definitions along each scale in which two definitions are bipolar and one definition is at the midpoint, and (c) having raters generate anchors unique to each instructor during the evaluation phase.
Chapter 2. Statement of the Problem

Identifying the critical properties which will improve the psychometric utility of an evaluation instrument requires a considerable expenditure of energy on the psychometrist's behalf (Borman & Dunnette, 1975). Unless this task is completed with diligence, the goal of psychometric soundness for the evaluation will not be met. An alternative solution to this problem would be to use an evaluation instrument developed elsewhere which has been shown to be effective. Borman and Vallon (1974) evaluated the utility of a BES form designed to measure the performance of administrators in a setting other than that in which the form was developed. The form was found to have limited generalizability. However, the BES form evaluated contained scales originally developed by Smith and Kendall (1963). Using these ten-year-old scales should decrease the effectiveness of the form. This leaves the possibility that a recent BES form developed in one setting may yield in another setting more favorable psychometric results than those of Borman and Vallon (1974).

This thesis will examine whether a recently developed BES form borrowed from its original setting has psychometric equivalence to a rating form developed specifically for a setting. In particular, the results of using three ratings
forms are examined: (a) a BES form developed in and for a Southeastern university setting, designated the Local form; (b) a BES form developed in and for the psychology department at the University of California--Berkeley, designated the Original form; and (c) the same BES form used at the University of California--Berkeley, but transferred for use at the Southeastern university, designated the Borrowed form. All three forms were designed to measure teaching effectiveness. Both the Local and Borrowed forms are compared in a relative sense in order to select the rating form which can best be utilized by the psychology department at the Southeastern university. All of the instruments are evaluated in an absolute sense in order to test the degree to which the instruments met basic psychometric requirements.

Comparisons are first made between the Original and Borrowed forms. Before the Borrowed form can be utilized for future evaluations, the ratings must have minimally desirable properties which are equal or superior to those of the Original. Bearing little resemblance to the Original would suggest that the Borrowed scales have ratings which are unstable. Generalizability to a new setting must be demonstrated before the Borrowed form can be considered psychometrically sound.

Second, comparisons are made between the Borrowed and Local forms. If properties of the Borrowed compared with the Local form are equally or more desirable, one can bypass the time consuming effort necessary for inhouse scale development. This assumes, of course, that the Borrowed scales have suc-
cessfully generalized to the new setting. Neither form will be deemed appropriate for making further evaluations if the Borrowed form has not generalized to the new setting even if it is equivalent or superior to the Local. On the other hand, inhouse scale development would be suggested if the Local form has substantially better psychometric properties than the Borrowed form.
Sample

Scale Development Phase. Forty undergraduate psychology students participated in developing the Local format. These forty students were placed into six groups, designated Groups A through F. Since attendance in the evaluation form development sessions was voluntary, the actual number of members in each group varied. The characteristics of the sample of developers are broken down by student year, developmental group, and sex in Table 1.

As indicated in Table 1, there were more underclassmen than upperclassmen participating in scale development, equal numbers of males and females participating, and more students participating in the later sessions than in the earlier sessions.

Scale Evaluation Phase. Once the Local form was developed, 23 undergraduate classes were administered either this form or the Berkeley form. The Local form was used in 12 classes, the Berkeley form in 11 classes, both forms being randomly assigned to underclassmen and upperclassmen courses.

Procedure

A series of conferences with all developers was held to
Table 1
The Sample of Developers for the Local Format

<table>
<thead>
<tr>
<th>Group</th>
<th>Freshmen</th>
<th>Sophomores</th>
<th>Juniors</th>
<th>Seniors</th>
<th>Σ</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Group A</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Male</td>
<td>2</td>
<td>1</td>
<td>2</td>
<td>0</td>
<td>5</td>
</tr>
<tr>
<td>Female</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td><strong>Group B</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Male</td>
<td>0</td>
<td>3</td>
<td>1</td>
<td>1</td>
<td>5</td>
</tr>
<tr>
<td>Female</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>2</td>
</tr>
<tr>
<td><strong>Group C</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Male</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>2</td>
</tr>
<tr>
<td>Female</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>3</td>
</tr>
<tr>
<td><strong>Group D</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Male</td>
<td>3</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>5</td>
</tr>
<tr>
<td>Female</td>
<td>2</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>3</td>
</tr>
<tr>
<td><strong>Group E</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Male</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>Female</td>
<td>2</td>
<td>1</td>
<td>2</td>
<td>0</td>
<td>5</td>
</tr>
<tr>
<td><strong>Group F</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Male</td>
<td>2</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>3</td>
</tr>
<tr>
<td>Female</td>
<td>4</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>6</td>
</tr>
</tbody>
</table>

| Σ          | 19       | 9          | 8       | 4       | 40 |

Groups labeled in chronological order (Group A worked first; Group F worked last)
stress the importance of producing a sound evaluation device. Each group was informed that the evaluation format should stress relevant teaching behaviors expected of an instructor within the psychology department. In addition, each group was provided with illustrations of the process of BES development as detailed by Smith and Kendall (1963). After providing each group with the major arguments concerning the advantages of proper scale development, each group was instructed to perform a particular task of scale development.

**Stage 1.** A conference was held with Group A developers in order to generate important teaching performance dimensions. Participants were required to define each performance dimension in as much detail as possible. Specifically, Group A developers were instructed to define each performance dimension generally, clarifying each area of performance by defining high, moderate, and low performance (clarification statements). As a result of this first meeting with Group A developers, nine performance dimensions were generated. A conference was then held with Group B developers. With the same essential task as Group A, Group B developers generated eight performance dimensions along with clarification statements. After their dimensions were generated, Group B developers were presented with the list of performance dimensions generated by Group A. Group B developers were thereupon instructed to consolidate the 17 dimensions generated from both conferences, eliminating overlapping and/or irrelevant dimen-
sions. After the consolidating process, nine performance dimensions along with clarification statements remained.

Stage 2. A conference was held with Group C developers in order to generate a list of potentially observable behavioral incidents that occur in the classroom. Areas of performance generated by developers of Group A and Group B were given to Group C participants. Group C developers revised dimensions and clarification statements that were considered ambiguous. After a group discussion, each individual wrote one behavior corresponding to each clarification statement across the nine dimensions. Group D developers had essentially the same task as Group C developers. During this stage 378 behavioral incidents were generated. Some incidents considered ambiguous by the author were rewritten or eliminated. After this editing process 211 items remained.

Stage 3. Group E developers were instructed to match the behaviors to their associated performance dimensions, that is, to retranslate the incidents to performance dimensions. Incidents not matched to the same performance dimension by a majority of raters were eliminated by the author. Dimensions not matched to at least three incidents were eliminated. After this process 155 incidents and all 9 dimensions remained.

Stage 4. A conference was held with Group F developers in order to rate the desirability of each behavioral incident on a scale from 1 (worst performance) to 7 (best performance).
After the raters judged the desirability of each incident, the mean and the standard deviation corresponding to each incident were calculated. The mean of the desirability ratings for an incident determines the point at which the behavior anchors the scale. The standard deviation reflects the degree of consensus among developers concerning the desirability of the incident. A high standard deviation reflects lack of consensus among student raters concerning the desirability of the particular behavioral incident. As in the original Smith and Kendall study (1963), those incidents associated with standard deviations greater than 1.5 were considered ambiguous and eliminated. Again, dimensions not matched to at least three incidents were eliminated.

Final Development. From the 69 incidents that met the criteria in all the previous stages, 37 were chosen by the author to anchor the nine dimensions. This final elimination was based on two criteria. First, behavioral incidents were eliminated if there was not adequate variability between the means corresponding to them, that is, along each scale's range the retained anchors were well spaced so that the raters would not find the dimensions ambiguous. Second, each dimension was structured so that each scale had at least three but not more than five anchors in addition to the clarification statements. The author assumed that too many anchors may be confusing to the rater during evaluation. This confusion may introduce unnecessary bias and error into the evaluation.
Table 2 shows the number of dimensions and behaviors remaining after each developmental phase and in the final development of the scale.

Table 2

Number of Dimensions and Behaviors Remaining After Each Developmental Phase of BES Format

<table>
<thead>
<tr>
<th>Dimension Description</th>
<th>Stage 2</th>
<th>Stage 3</th>
<th>Stage 4</th>
<th>Final Development</th>
</tr>
</thead>
<tbody>
<tr>
<td>Concern for the Student</td>
<td>28</td>
<td>23</td>
<td>8</td>
<td>5</td>
</tr>
<tr>
<td>Class Interest</td>
<td>25</td>
<td>17</td>
<td>7</td>
<td>4</td>
</tr>
<tr>
<td>Relations with Students</td>
<td>22</td>
<td>16</td>
<td>8</td>
<td>3</td>
</tr>
<tr>
<td>Instructor Attendance</td>
<td>21</td>
<td>18</td>
<td>6</td>
<td>4</td>
</tr>
<tr>
<td>Class Participation</td>
<td>30</td>
<td>26</td>
<td>8</td>
<td>4</td>
</tr>
<tr>
<td>Motivation of Instructor</td>
<td>21</td>
<td>12</td>
<td>8</td>
<td>4</td>
</tr>
<tr>
<td>Sensitivity to Class Progress</td>
<td>17</td>
<td>14</td>
<td>7</td>
<td>5</td>
</tr>
<tr>
<td>Grading</td>
<td>19</td>
<td>17</td>
<td>5</td>
<td>5</td>
</tr>
<tr>
<td>Teacher Preparation</td>
<td>18</td>
<td>12</td>
<td>12</td>
<td>3</td>
</tr>
<tr>
<td><strong>Total Behaviors</strong></td>
<td><strong>211</strong></td>
<td><strong>150</strong></td>
<td><strong>69</strong></td>
<td><strong>37</strong></td>
</tr>
</tbody>
</table>

As indicated in Table 2, initially each dimension was matched with at least 17 behaviors. At each developmental phase, several incidents were lost; the greatest loss occurred during the desirability rating phase (Stage 4).
Developmental Characteristics of Berkeley and Local Forms. Both the Local form and the Berkeley form have a BES format with common structural characteristics. Table 3 shows the areas of performance for both the Local and the Berkeley forms. As indicated in Table 3, both forms have the same number of dimensions. Each dimension is scaled from 1 to 7 and responses are whole numbers only. Yet, both forms have unique structural variations. First, the Berkeley form has a greater variety of behavioral anchors than the Local form. Moreover, the Berkeley form anchors are much longer and more specific to class situations than are those of the Local form. Since the present author assumed that longer behavioral anchors would be too confusing to the student raters, only shorter behavioral descriptions were chosen during the editing process to anchor each scale. Second, only the Local form contained clarification statements along each scale. Finally, the Local form contained an open-ended question designed to gather unique behaviors of teaching performance not necessarily indicated within the form. Responses to this question can provide a pool of behaviors which can be used as potential anchors in further refinements of the BES form. The Berkeley form had neither the clarification statements nor a question designed to gather further behaviors.

Administration. Instructors were identified by special codes to protect their anonymity during the analysis of the responses. Since the Berkeley and the Local forms were both
### Table 3
Comparison of the Dimensions and the Number of Anchors Matched to Each, of the Berkeley Form and the Local Form

<table>
<thead>
<tr>
<th>Berkeley Form</th>
<th>Local Form</th>
</tr>
</thead>
<tbody>
<tr>
<td>Dimension</td>
<td>Anchors</td>
</tr>
<tr>
<td>Ability to Motivate Students</td>
<td>10</td>
</tr>
<tr>
<td>Relevance</td>
<td>8</td>
</tr>
<tr>
<td>Interpersonal Relations</td>
<td>10</td>
</tr>
<tr>
<td>Testing</td>
<td>8</td>
</tr>
<tr>
<td>Organization</td>
<td>8</td>
</tr>
<tr>
<td>Work Load</td>
<td>8</td>
</tr>
<tr>
<td>Delivery</td>
<td>9</td>
</tr>
<tr>
<td>Grading</td>
<td>8</td>
</tr>
<tr>
<td>Depth of Knowledge</td>
<td>9</td>
</tr>
</tbody>
</table>

<sup>a</sup>Clarification statements are included in the count of anchors for the Local form.
administered at the end of the semester, students were very familiar with the instructor’s performance.

Data Analysis

In order to provide a check for bias and error between formats, four dependent variables were used to analyze the student evaluations. Each variable serves as an index of a format’s relative usefulness within the criterion situation.

Leniency Effects. A format will lack leniency error if the average of all the ratings of all instructors by all student raters is at the midpoint of the scale. Leniency was determined by finding the mean across all students and instructors for each performance dimension. Using the performance dimensions’ means as data points, a t test between formats was calculated in order to determine if a significant mean difference favored either format. Since both the Berkeley form and the Local form were scaled to require responses from 1 through 7, any dimensional mean substantially higher or lower than 4, the midpoint, reflects leniency error. The format having dimensions whose means least differ from their midpoints has the greater psychometric utility (Bernardin, LaShells, Smith, & Alvares, 1976).

Discriminability. A format possesses discriminability when there is a concensus among student raters concerning the quality of performance exhibited on each dimension. Discriminability was calculated by finding the variance of desirability ratings by student raters for each instructor on each per-
formance dimension. A substantial amount of variance ($\sigma \geq 1.5$) shows a lack of consensus among students concerning the quality of an instructor's performance on a dimension. Standard deviations across instructors were then averaged for each performance dimension. A $t$ test was calculated between formats in order to determine if there were significant differences in standard deviations on dimensions. The format having smaller standard deviations has the greater psychometric utility (Bernardin, LaShells, Smith, & Alvares, 1976).

Variability. A format will possess variability, that is, discriminant validity (Campbell & Fiske, 1959), when students evaluate instructors on specific performance dimensions rather than in an overall global fashion. Variability was determined by correlating the rated responses to each performance dimension with the rated responses to every other performance dimension across students for an instructor. Stated differently, a matrix was formed by correlating the ratings on one dimension with the ratings on all other dimensions for each instructor. A mean for each correlation matrix was found by transforming all correlations to Fisher's $Z$'s and then averaging the $Z$'s (Bernardin, LaShells, Smith, & Alvares, 1976). A low mean correlation suggests that the instructor was rated on each performance dimension independently. A high mean correlation suggests that the instructor was rated in a global fashion (contains halo). Using the instructors' mean correlations as data points, a $t$ test for independent measures was
calculated in order to determine if there was a significant difference in variability between formats. The format having a lower mean correlation across all instructors has less biased ratings.

**Interrater reliability.** A format contains interrater reliability when there is agreement among student raters across dimensions when evaluating an instructor's performance. Interrater reliability was calculated by correlating the ratings by dimensions of each student with every other student for each instructor (Bernardin, LaShells, Smith, & Alvares, 1976). A mean for each correlation matrix was calculated by transforming all correlations to Fischer's $Z$'s and then averaging the $Z$'s. A high mean for an instructor suggests that the students generally agree about the quality of the instructor's performance. A low mean for an instructor suggests a lack of agreement about the instructor's performance in class. Using the instructor's means as data points, a $t$ test for independent measures was again calculated in order to determine if there were significant mean differences in reliability between formats (Bernardin, LaShells, Smith, & Alvares, 1976). The format having mean correlations which are higher across instructors has better reliability.
Chapter 4. Results

Bias and error are ubiquitous problems in performance evaluation. To provide a check for bias and error, not only should forms be compared on relative terms, but the absolute soundness of each form should be analyzed also. Although one format may be relatively better than another, both may be unsound and not useful within the criterion situation.

Effects of Leniency

Leniency error is the first criterion used to test the soundness of evaluations. Leniency reflects the overall generosity of student ratings given across all instructors. Table 4 shows empirical comparisons between the Berkeley form and the Local form concerning the extent to which dimensional means are subject to leniency error. For both formats, each dimension's mean is above its midpoint. This finding suggests that both formats are subject to leniency error. A t test for independent measures indicates a significant mean difference in favor of the Berkeley format, \( t(8) = 17.62, p < .01 \). This finding suggests that the Berkeley format yields ratings which are substantially less lenient than the Local format. Even though relatively better than the Local, the Berkeley form does not have absolute soundness. Mean ratings which drastically depart from the midpoints of the
### Table 4
Comparison of the Means and Standard Deviations of the Dimensions of the Berkeley Form and the Local Form

<table>
<thead>
<tr>
<th>Dimension</th>
<th>X</th>
<th>S.D.</th>
<th>Dimension</th>
<th>X</th>
<th>S.D.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ability to Motivate Students</td>
<td>5.34</td>
<td>1.94</td>
<td>Concern for the Students</td>
<td>6.05</td>
<td>.99</td>
</tr>
<tr>
<td>Relevance</td>
<td>5.71</td>
<td>1.38</td>
<td>Class Interest</td>
<td>5.16</td>
<td>1.49</td>
</tr>
<tr>
<td>Interpersonal Relations</td>
<td>6.10</td>
<td>1.26</td>
<td>Relations with Students</td>
<td>5.89</td>
<td>1.31</td>
</tr>
<tr>
<td>Testing</td>
<td>4.71</td>
<td>1.18</td>
<td>Instructor Attendance</td>
<td>6.22</td>
<td>1.26</td>
</tr>
<tr>
<td>Organization</td>
<td>5.46</td>
<td>1.07</td>
<td>Class Participation</td>
<td>5.18</td>
<td>1.13</td>
</tr>
<tr>
<td>Work Load</td>
<td>5.49</td>
<td>1.39</td>
<td>Motivation of Instructor</td>
<td>6.13</td>
<td>1.41</td>
</tr>
<tr>
<td>Delivery</td>
<td>5.32</td>
<td>1.56</td>
<td>Sensitivity to Class Progress</td>
<td>5.88</td>
<td>1.39</td>
</tr>
<tr>
<td>Grading</td>
<td>6.22</td>
<td>1.22</td>
<td>Grading</td>
<td>6.25</td>
<td>1.14</td>
</tr>
<tr>
<td>Depth of Knowledge</td>
<td>5.63</td>
<td>1.32</td>
<td>Teacher Preparation</td>
<td>6.06</td>
<td>1.14</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>X</th>
<th>5.47</th>
<th>1.45</th>
</tr>
</thead>
</table>

| X   | 5.97 | 1.22 |
scales suggest that the Berkeley form as well as the Local form is subject to leniency error.

**Effects of Discriminability**

A format possesses discriminability if there is a consensus among students concerning an instructor's quality of performance on each dimension; this consensus is indicated by low standard deviations on dimensions. Table 4 shows the standard deviations for each dimension for both formats. As indicated in Table 4, two Berkeley-format dimensions, Ability to Motivate Students and Delivery, have standard deviations exceeding 1.5. This finding suggests that student raters could not arrive at a consensus on these two performance dimensions of the Berkeley format. As further indicated in Table 4, no Local-format dimension has a standard deviation exceeding 1.5. A t test for independent measures indicates that there are no significant differences between formats, \( t(8) = 1.03, p > .05 \). This finding suggests that consensus among raters concerning the quality of instructors' performances does not differ for the two forms.

**Effects of Variability**

Variability is the third criterion used to test the soundness of evaluations. Variability concerns the degree to which students evaluated instructors independently on each dimension rather than in an overall global fashion. Table 5 shows summary statistics comparing the forms on variability.
Table 5

Comparison of Summary Statistics Indicating the Relative Soundness of Student Ratings on the Berkeley and Local Forms

<table>
<thead>
<tr>
<th>Dependent Variable</th>
<th>Berkeley Form</th>
<th>Local Form</th>
</tr>
</thead>
<tbody>
<tr>
<td>Leniency Error</td>
<td>5.47</td>
<td>5.97</td>
</tr>
<tr>
<td>Discriminability</td>
<td>1.45</td>
<td>1.22</td>
</tr>
<tr>
<td>Variability</td>
<td>.45</td>
<td>.30</td>
</tr>
<tr>
<td>Interrater Reliability</td>
<td>.15</td>
<td>.20</td>
</tr>
</tbody>
</table>

As indicated in Table 5, on variability both the Berkeley and the Local formats had mean correlations at or exceeding .30. A mean correlation of .30 does not reach statistical significance (p > .05). This finding suggests that both forms satisfied the variability criterion, that is, each form appeared to distinctly measure reasonably independent performance dimensions. A t test for independent measures revealed a significant mean difference favoring the Local form, t (10) = 1.94, p < .05. Stated differently, students tended to rate instructors the same across performance dimensions more often when using the Berkeley than the Local form.

Effects of Interrater Reliability

Interrater reliability was perhaps the most critical test in the evaluation. A form possessing interrater reliability suggests that raters agreed with one another regarding
the instructor's performance across the dimensions. Interrater reliability therefore limits the form's validity. As indicated in Table 5, neither form had a mean interrater reliability exceeding .20. A form possessing adequate interrater reliability should have a mean coefficient exceeding .60 (Nunnally, 1967, p. 227). This finding suggests that neither form produced rater consistency in the assessment of instructors' performances. Moreover, a $t$ test for independent measures indicates that there were no significant differences between forms in the consistency between raters, $t(10) = 1.18$, $p > .05$. 
Chapter 5. Discussion

Whenever performance is evaluated by a group of raters, an assumption is made that the resultant evaluations index the effectiveness of the performer. Before student ratings can be accepted as useful accounts about an instructor's performance in class, standards suggesting that the forms are sound must be satisfied. Both absolute and relative standards were established in order to determine if the Berkeley form and the Local form index teaching effectiveness.

Evaluation of Scales in an Absolute Sense

Determining if evaluations are sound in absolute terms is critical. Standards which focus upon absolute soundness lay the framework upon which performance ratings can be scientifically assessed. A systematic framework applied across several rating devices will show if advancements are being made toward consistently rating human performance. Ultimately, achieving absolute soundness would suggest that performance can be impartially rated by humans. Moreover, an evaluation form would be selected which reflects useful accounts of teaching effectiveness. Since absolute soundness has not been met totally by any rating device, critical properties within the form or tendencies within the raters must be identified which favor sound ratings. Multiple measures of abso-
lute soundness were established to evaluate the soundness of the Local and the Berkeley forms.

Ratings are unsound if the evaluations possess leniency error. Having mean ratings near the midpoint of the scales suggests that the evaluations are sound in an absolute sense. Both the Local and the Berkeley forms had mean ratings across all dimensions which were above the respective midpoints. Regardless of the form used, students were generous in their ratings of instructors. This made it difficult to identify specific classroom teaching deficiencies and strengths. In other words, lenient ratings made it difficult to provide meaningful feedback to the instructor. Moreover, lenient ratings created a difficulty in distinguishing between instructors who were outstanding in performance and those who were not. Thus, neither the Berkeley nor the Local form met the first requirement of absolute soundness since the ratings were subject to leniency error.

Ratings are sound if the evaluations possess discriminability. Evaluations will meet this requirement if students can arrive at a consensus concerning the performance of an instructor. Ratings which are dispersed from the mean more than 1.5 standard deviations lack discriminability and thus do not meet the second requirement for absolute soundness. For the Local form, all performance dimensions possessed discriminability. Whenever students were rating an instructor on a performance dimension, they tended to arrive at a
consensus concerning the instructor's performance. However, for the Berkeley form, two dimensions did not meet the requirement of having ratings with small dispersions from the mean. For the other seven performance dimensions, students arrived at a consensus. Thus overall, both the Berkeley and the Local forms approximated the requirements of having absolute soundness inasmuch as the students agreed about the performance of an instructor on a performance dimension.

Third, ratings have absolute soundness if they contain variability. Performance dimension evaluations are distinct if the dimensions are uncorrelated. For the Local form, this requirement was easily met. Low associations among performance dimension evaluations suggest that the ratings assigned by students in one performance area only slightly influenced the ratings given in other areas. Yet, the lack of agreement between raters concerning the performance of instructors, i.e., interrater reliability, may account for the low inter-dimensional correlations. On the other hand, the correlations among the Berkeley form dimensions approximated statistical significance, the point at which the form would be judged to lack variability. Several instructors were rated the same across several performance dimensions on the Berkeley form, obscuring intra-individual differences of teaching behaviors (Helmstadter, 1964, p. 191). In other words, the ratings on one performance dimension influenced the ratings on other dimensions. Thus, the absolute soundness require-
ment of variability was easily met by the Local form but was only minimally met by the Berkeley form.

Finally, ratings have absolute soundness if the evaluations are reliable measures of performance. When forms are first being developed, Nunnally (1967) argues that the ratings should have reliability coefficients of around .67. The ratings on both the Local and the Berkeley forms had reliability coefficients of about .20. This finding suggests that students were inconsistent in their ratings. Regardless of format, ratings given by one student in no way coincided with ratings given by other students. Ratings such as these indicate that students either had difficulty understanding the forms or could not agree on the desirability of the instructors' performances in class.

In summary, both the Berkeley and the Local forms met two of the four standards for absolute soundness, at which all evaluations should aim. Specifically, both forms possessed discriminability and variability.

**Evaluation of Scales in a Relative Sense**

Relative psychometric comparisons between scales serve to identify the format more reflective of sound ratings. Within the present study, both the Berkeley and the Local forms possessed some psychometric properties favoring one form over the other.

First, the Berkeley form evaluations were significantly less lenient than the Local. This finding suggests that
providing meaningful feedback to an instructor about performance in class was easier with use of the Berkeley form than with use of the Local form. Mean ratings nearer the scales' midpoint, 4, indicate that students using the Berkeley form tended to differentiate effective instructors (those with ratings above the scales' midpoint) from ineffective instructors (those with ratings below the scales' midpoint) across performance dimensions. Higher mean ratings on the Local form were probably due to anchoring behaviors to which the students could not relate their perceptions of teaching performance. Compared with the Berkeley form, the Local form had fewer and less specific behavior descriptions anchoring each scale. Keveane: and MacGann (1975) found that evaluations using forms containing highly descriptive statements to anchor the scales had fewer lenient ratings than evaluations using forms without behavior descriptions to anchor the scales. This finding was corroborated by the findings of the present study. Thus, the Local form needs to be revised in order to include better anchors.

Second, the Local form was equivalent to the Berkeley form with regard to discriminability. Regardless of the form used, students' ratings of instructors' performances did not deviate very much from the average.

Third, the Local form evaluations were significantly more variable than the Berkeley. This finding suggests that locally generated dimensions were more clearly distinguished by the students. Borman and Vallon (1974) found that, when
the areas of performance and behavioral descriptions were
developed elsewhere, the rating on one performance dimension
influenced the ratings on other dimensions. This finding is
consistent with the results of the present study. The Berkeley
form, borrowed from its original setting for use in the
present study, was more subject to halo error than was the
Local form. Thus, indigenous intensive development of scales
may be a way of increasing variability and reducing halo
error.

Finally, the Local form was equivalent to the Berkeley
form with regard to reliability. Since both forms had unreli-
able ratings, those ratings also tended to have poor validity
(Helmstadter, 1964, p. 84).

In summary, the Berkeley form evaluations were less leni-
ent but were not as variable as the Local form evaluations.
Participation in scale development by the student raters and
use of several descriptive incidents as anchors are critical
properties for improving the psychometric utility of these
scales.

Comparison between the Original Berkeley Form and
the Borrowed Berkeley Form

Poor generalizability from the original study may
account for the unfavorable results obtained using the Berkeley
form in this study. Comparing the Berkeley form evalua-
tions originally obtained by Zedeck, Jacobs, and Kafry (1976)
with those obtained in the present study, the Borrowed Berke-
ley ratings were more lenient and had greater discriminability. Inasmuch as measures of variability were not determined for the Original Berkeley, this finding suggests that the Berkeley may not be designed to focus upon the unique differences in an instructor's performance. If so, the form as originally developed may have been subject to halo error. Since the measure of reliability in the original setting is unknown, the Berkeley form may have been unstable for the original study as well as for this study.

In summary, the measures of leniency and discriminability suggest that the Berkeley did not generalize well into a different setting.

Comparison between the Local Form and the Original Berkeley Form

Compared with the Original Berkeley form developed by Zedeck, Jacobs, and Kafry (1976), the Local form yielded ratings which were more lenient but which had greater discriminability. The original study by Zedeck, Jacobs, and Kafry gave no measure of variability, a serious omission--from the point of view of this author--since greater variability or lack of halo error was a principal advantage of the Local form evaluations over those of the Borrowed Berkeley form.

Compared with the Original Berkeley, the Local form had greater discriminability and was more lenient. These findings suggest that the Local form did not have the utility evidenced in the Berkeley form as originally developed.
Implications of Results

In view of the psychometric considerations presented in this thesis, the results offer little support for the use of a BES format for judging teaching performance. Three major implications can be drawn from this study.

First, caution must be exercised when a BES evaluation form is transferred from its original setting to a new one. According to BES format rationale, developers should generate a set of performance dimensions and anchoring behavioral descriptions with which the raters can identify. Difficulty in understanding language within the Original Berkeley form may have been a major problem, limiting its effectiveness as a Borrowed form used in a different setting. Also, ambiguity of the Original Berkeley form when used in the new setting may have been due to anchors not considered germane by raters in the new setting.

Second, intensive inhouse development of a BES form did not produce sound psychometric ratings. Even though the Berkeley form did not generalize to the new setting, there was no evidence that the Local form was better than the Borrowed. Ratings from both forms gave evidence of unfavorable psychometric characteristics. A major problem with the Local form may have been improper editing. Alterations made by the author may have affected the scales' utility. Anchoring behavioral descriptions were shortened and limited to not more than five along each scale. Such alterations may have
made the form ambiguous. Proper development of a BES form, therefore, requires considerable care on the part of the psychometrist so that error will not be introduced into the ratings.

Finally, the failure of the Berkeley scale to retain its psychometric properties when employed locally may suggest that BES scales are inherently non-transportable. However, this conclusion would appear premature given the construction of the present investigation. While superficially one might contend that the two rating situations were similar insofar as both were carried out in a university milieu, closer examination reveals substantial differences in both ratee and ratee characteristics.

Ratee or instructor differences between the two institutions may reduce the content validity of both the dimensions and their anchors. Instructors at Berkeley typically are engaged in research as a primary activity. Teaching generally receives less emphasis both in terms of its impact on the instructor's total performance and the amount of time spent in the classroom. Generally, Berkeley instructors teach one or two classes a semester contrasted with four at the "local" university. Moreover, Berkeley instructors are more likely to teach courses only in their specialty, whereas "local" instructors teach a wider diversity of courses. In short, the total job descriptions are substantially different for the instructors in the two universities. More comparable job
descriptions might have produced more comparable scale characteristics.

Rater or student differences, which were also substantial, are even more likely to have had an impact on the scale. While the "local" university draws its students primarily from Kentucky and its contiguous states, Berkeley students are typically Californians. The academic credentials of the entering students are highly discrepant (Furniss, 1973). The qualitative differences in the elementary and secondary school systems in California vis-à-vis Kentucky require little elaboration. Admission requirements are considerably more stringent at Berkeley than at the "local" university. Specifically, Berkeley requires a B average or better and selects only 76% of its applicants. On the other hand, the "local" university requires a C average or better and accepts 94% of its applicants.

In addition to population differences at the time of admission, composition of the undergraduate school bodies is discrepant. A greater percentage of undergraduate students receive the baccalaureate degree at the end of the academic year at Berkeley (30%) than at the "local" university (17%).

These large differences between raters at the two institutions suggest that the appropriateness of the Berkeley scale in the "local" setting is suspect. The language used in the Berkeley scale may well have been unfamiliar to raters in the "local" setting. Moreover, one might anticipate
Berkeley raters to be more cognitively complex, making finer discriminations generally than do the "local" raters.

Thus, despite the apparent similarities between the two contexts, large rater differences may render any common scale inappropriate. Vigilance must be exercised in assessing rater similarities and differences when "borrowing" a scale developed elsewhere.
Appendix

The following papers are photographic reproductions, reduced in size, of the performance dimension scales as used in the evaluation setting.
Dear Psychology Student:

The following is an instructor/course evaluation form. It has been developed by students taking psychology courses both here at Western and at the University of California-Berkeley.

Eighteen scales are included in this rating form. You will note that each scale has the following components: 1) a one-line definition of what the particular scale is designed to measure, 2) a one (bad performance) to seven (good performance) rating scale, and 3) a list of various behaviors along each scale.

The behaviors listed to the right of each scale are there only to help clarify the meaning of the seven numbers along each scale. They are simply examples of the kinds of behaviors that might be expected from an instructor receiving a rating at various points along the scale. Your instructor need not actually have engaged in one of the particular behaviors for you to check that point on the scale. But based on your experience this semester, you have an idea which one of the behaviors represents the type of action to be expected from your instructor.

Your task is to decide which one of the seven numbers on each scale best describes the level of performance of your instructor. Then blacken the corresponding number for that scale on the IBM answer sheet. Please do not mark in the booklet. The scales are numbered one through eighteen. Therefore, when finished, you should have blackened one space for each of the first eighteen items on the answer sheet. Please note that although the answer sheet has response options from 0 through 9, all your responses to the eighteen scales should fall within the response option 1 through 7.

Once again, you are to respond once to each scale by blackening the corresponding number on the answer sheet.

Your instructor and the Psychology Department sincerely appreciate your thoughtful completion of this scale. We will make every effort to actively use the information you provide to improve the quality of psychology courses and your instructor's teaching skills.
ABILITY TO MOTIVATE STUDENT--the professor's generation of students' interest in the subject matter.

This professor could be expected to be so inspiring that the student is often ahead in his reading assignments.

After completing an introductory social psychology course with this professor, most students could be expected to enroll in other classes that deal with the field of social psychology.

In this developmental psychology class, if a student hesitantly describes a little experiment with school children that he is thinking about, this professor could be expected to reply: "Great! It sounds good. Your plan has some flaws, but every psychologist's plan has some flaws at first. We can work it out, and I'm sure you'll enjoy doing it!"

In an introductory psychology class, this professor could often be expected to pose questions and issues to students that are later discussed in section meetings or with classmates and friends outside of class.

This professor's students could be expected to have no qualms about studying the material he assigns.

--The students in this professor's class could be expected to do the required work.

--The students in this professor's class could be expected to do the required work and no more.

--This professor of a psychological statistics class could be expected to try to push students into being interested by almost pleading with them.

--Attendance in this professor's class could be expected to be less than 50% each meeting.

After completing an introductory psychology course with this professor, most students could be expected to be so disillusioned with psychology that they have little desire to enroll in other psychology courses.
DELIVERY--the professor's ability and way of conveying the material.

This professor could be expected to have a clear, distinct, excellent voice and can be heard anywhere in the auditorium. He could be expected to speak with inflection and to convey each mood of the material.

This professor's use of visual aids could be expected to entertain and inform the students.

This professor, when contrasting operant and classical conditioning could be expected to make good use of blackboard.

This professor's voice could be expected to be clear and distinct but sometimes he could be expected to speak too fast for the student to get the material into his notes.

In this introductory psychology class, students could be expected to have no difficulty understanding this professor's lecture on conditioned-response and response sets, but they could often be expected to be bewildered when he discourses on theory in general.

When lecturing, this professor could be expected to pace across the platform back and forth and make the students nervous.

On occasion, this professor could be expected to mumble to himself in the middle of a lecture.

In order to study for an exam of this professor's, students could be expected to go to the TAs because they can't understand the explanations of the professor.

This professor could be expected to read from his notes and to speak in a low monotone. It is almost impossible not to become drowsy during class.
DEPTH OF KNOWLEDGE--the professor's mastery of the subject matter.

1. In a statistics class, instead of admitting that he does not know the answer to a question about multiple regression, the professor could be expected to offer a reply in vague general terms that confuses the students even more.

2. If a student asks this professor a question, the student often could be expected to feel that the professor--in a very round-about way--is merely feeding the student back his question.

3. This professor of experimental psychology could be expected to know the material about verbal learning and conditioning that is covered in the reading assignments but beyond that does not elaborate.

4. If a student asks this professor of a class on motivation to elaborate on Maslow's theory, the professor could be expected to provide a general outline but must refer the student to other sources for specifics.

5. If a student happened to read an article about classical conditioning of the octopus and asked his professor of introductory psychology for more details, the professor could be expected to say, "This is not my area," but would give the student references for finding more information on classical conditioning of animals.

6. This professor of comparative psychology when asked about where to look for material on the social behavior of gorillas could be expected to give the student a dozen names of books and their authors, as well as many other parts of the literature to look into.

7. This professor of a personality course when discussing Adler and Rogers could be expected to handle a question that came up during lecture that isn't covered satisfactorily in assigned readings.

This professor of a personality course knows the history of the subject matter so thoroughly, even to the minor details, that he could be expected to sort out the trivia from the important parts and present the important parts in a very simple manner.
GRADING— the ways and purposes for which the professor uses grades.

6. This professor could be expected to allow each student options; e.g., papers, projects, and/or final exam, upon which the student’s course grade will be based.
   This professor could be expected to mark off for papers which are a few days late; and when grading the exam, if the student is near the "borderline," he could be expected always to receive the higher grade.
   In this professor’s class, any suggested assignments or student’s own projects could be expected to be extra credit and can only help the student’s grade (which is based on exams).
   The course grade could be expected to be based on three 4-page papers on critical topics covered during the course.

5. This professor could be expected to drop a student’s grade on a paper from an "A" to a "B" because the student hands in the paper a week late.
   If a student with a poor statistical background and ability enrolls in a required psychology statistics class and gets a "D" on the first test but then earns a "B-" on the next exam and an "A" in the final exam, this professor could be expected to not consider giving the student an "A" for the course because the student’s "average" grade is less than "A"
   Each student in this professor’s developmental psychology class works on a 5-week project, and the professor could be expected to rigidly adhere to his system where the project grade is lowered one letter grade for each day it is turned in after the deadline.
   This professor could be expected to give out only so many "A’s," so many "B’s," so many "C’s," etc; there is a predetermined number of students for each grade.

4. This professor could be expected to allow each student options; e.g., papers, projects, and/or final exam, upon which the student’s course grade will be based.
   This professor could be expected to mark off for papers which are a few days late; and when grading the exam, if the student is near the "borderline," he could be expected always to receive the higher grade.
   In this professor’s class, any suggested assignments or student’s own projects could be expected to be extra credit and can only help the student’s grade (which is based on exams).
   The course grade could be expected to be based on three 4-page papers on critical topics covered during the course.

3. This professor could be expected to drop a student’s grade on a paper from an "A" to a "B" because the student hands in the paper a week late.
   If a student with a poor statistical background and ability enrolls in a required psychology statistics class and gets a "D" on the first test but then earns a "B-" on the next exam and an "A" in the final exam, this professor could be expected to not consider giving the student an "A" for the course because the student’s "average" grade is less than "A"
   Each student in this professor’s developmental psychology class works on a 5-week project, and the professor could be expected to rigidly adhere to his system where the project grade is lowered one letter grade for each day it is turned in after the deadline.
   This professor could be expected to give out only so many "A’s," so many "B’s," so many "C’s," etc; there is a predetermined number of students for each grade.

2. This professor could be expected to allow each student options; e.g., papers, projects, and/or final exam, upon which the student’s course grade will be based.
   This professor could be expected to mark off for papers which are a few days late; and when grading the exam, if the student is near the "borderline," he could be expected always to receive the higher grade.
   In this professor’s class, any suggested assignments or student’s own projects could be expected to be extra credit and can only help the student’s grade (which is based on exams).
   The course grade could be expected to be based on three 4-page papers on critical topics covered during the course.

1. This professor could be expected to allow each student options; e.g., papers, projects, and/or final exam, upon which the student’s course grade will be based.
   This professor could be expected to mark off for papers which are a few days late; and when grading the exam, if the student is near the "borderline," he could be expected always to receive the higher grade.
   In this professor’s class, any suggested assignments or student’s own projects could be expected to be extra credit and can only help the student’s grade (which is based on exams).
   The course grade could be expected to be based on three 4-page papers on critical topics covered during the course.
INTERPERSONAL RELATIONS WITH STUDENTS--the professor's rapport with and sensitivity to students.

1. When a class doesn't understand a certain concept or feels "lost," this professor could be expected to sense it and act to correct the situation.

2. This professor could be expected to answer the student's questions about learning and conditioning without making the student feel stupid and without making the student feel that he's bothering the professor.

3. When confronted with questions after class, this statistics professor could be expected to stay and talk to the students until the next class must begin.

4. This professor, when a student comes to his office for help, could be expected to go through one explanation of the material and tell the student to read certain chapters of the text and to come back if he still has troubles understanding the material.

5. During lectures, this professor could often be expected to tell students with questions to see him during his office hours.

6. If a student asks this statistics professor to help him with "t"-tables a few days before the final exam, this professor could be expected to say that he has no time because he is very busy composing the exam, and to tell the student to ask a TA.

7. This professor could be expected to not see students individually, except during his regularly scheduled office hours.

8. This professor is never in his "official office." He could be expected to maintain his office in another part of the campus where he does his research and in order to learn of its whereabouts, students must ask him individually.

9. In this experimental psychology class, if a student approaches this professor after a lecture on visual-search and tells the professor that he is interested in devising an apparatus that will measure visual-search time more efficiently than present methods, the professor's attitude could be expected to be an "I-really-don't-care-if-you-do-it-or-not."

10. This professor could be expected to try to humiliate or embarrass students who disagree with him.
This professor of experimental psychology could be expected to integrate the reading material and the laboratory work with his lectures.

In an experimental psychology class, this professor, if intending to lecture on reaction time and its measurement, could be expected to have all the necessary apparatus set up before class.

This professor's lectures could be expected to pick up where the last one fell off.

This professor of history of psychology could be expected to organize his lectures so as to cover psychologists and movements in the field of psychology in chronological order.

This professor could be expected to get sidetracked at least once a week in lecture and, thereby, not cover material he has intended to.

This professor's schedule could be expected to leave him trying to teach the structure of "intelligence" without necessary background material being covered.

This professor of introductory psychology could be expected to assign readings on verbal learning and memory while his lectures and the section meetings are devoted to the study of power and authority.

This professor could be expected to tell the class to read Chapters 3, 4, and 5 and then lecture about material in Chapters 7, 8, and 9.
When this professor of social psychology lectures on social norms and role playing, he could be expected to give students vivid examples of how they as students and he as a faculty member play role games and how they have been socialized in student norms and faculty norms.

This professor could be expected to devote time from his planned lecture on power and authority, if during the lecture students become interested in discussing the psychological effects of power and authority in the Armed Forces.

This introductory psychology professor when discussing attitude change and opinion formation, could be expected to relate the lectures to specific actions taken by interest groups such as the John Birch Society, the National Rifle Association, or oil lobbies.

In a developmental psychology class, this professor could be expected to gear the course as to how students can raise their children.

This professor could be expected to be more interested in having the students learn the mechanics of theories of "achievement" rather than discussing their current status in terms of scientific and political implications.

This professor of social psychology, when discussing "conformity" and "deviant from social norms" never could be expected to refer to the "generation gap," new styles in clothing, new thinking about morality, or "street people" and "hippies."

This professor of psychological testing could be expected to try to avoid or discuss specific questions relating to the ethics involved in testing minority groups.

This professor of a class on motivation, even when specifically asked to relate some of his lectures to human beings instead of white rats could be expected to ignore this request and not mention motivation of humans once during the quarter.
TESTING—the ways and purposes for which the professor uses tests.

6. This professor of social psychology could be expected to give a test on cognitive balance theory such that the students often feel that they have learned something new about balance theory just from taking the test.

If this professor of personality gives an exam on the theories of Carl Rogers, it could be expected to be a short-essay exam in which the student uses material from many sources (lecture, textbook, outside readings, and personal experiences).

This professor could be expected to give two midterms and a final exam, each one consisting of half multiple choice and half essay.

This professor of introductory psychology could be expected to give multiple choice exams which ask only for specific facts which can be gotten directly out of the text or notes.

This physiological psychology professor could be expected to give multiple choice tests which require students to recognize verbatim statements of the assigned chapters. This professor's exams could be expected to be all True-False questions. The students who perform best on this professor's tests could be expected to be those who can memorize the material the most.

Students could often be expected to say about this professor: "his test items are so ambiguous" or "those test questions are really tricky."

This professor's exams could be expected to often stress material that has been briefly or lightly covered in class and to often devote little space to material which has been emphasized in class.
ASSIGNMENTS AND WORKLOAD—quantity and quality of the class requirements.

1. In this social psychology class, this professor's reading assignments could be expected to be large enough so that the student really feels that he is covering the field of social psychology but small enough so that it is practical to be enrolled in 3 other classes at the same time.

2. This professor, in addition to a regular reading list could be expected to hand out an optional reading list and to assure students that they are not going to be tested on the optional readings.

3. When this professor's psychological statistics class discusses computer programming, he could be expected to assign each student a program to be completed in 2 weeks, but he allows enough computer time so that each student will have ample time for rerunning and correcting errors in input.

4. In this psychological statistics class, this professor could be expected to require 3 hours of problem-solving activity per week, but 2 hours of this consists of section meetings with TAs helping and only 1 hour homework at home.

5. In this physiological psychology class, this professor could be expected to assign from 1 to 2 chapters of the text per week and a chapter of a lab manual before each laboratory session.

6. In this experimental psychology class, this professor could be expected to assign (in addition to regular readings from the text) one experiment before each class on Monday, Wednesday, and Friday.

7. In this industrial psychology class, this professor's reading assignments could be expected to be sporadic; one week it is a 25-page chapter from the book and the next week it may be two 30-page chapters and five journal articles.

8. In this psychological statistics class, this professor could be expected to not only assign from 20 to 30 problems a week (about 5 hours' worth of work) but, to also insist on covering one chapter of the textbook per week, regardless of the chapter's length or difficulty and regardless of whether students really understand the previous chapters.
CONCERN FOR THE STUDENT - the instructor encourages the student to seek help when he needs it.

Instructor is always available for the student during office hours and class.

The instructor makes himself available by giving office hours, phone number, as well as time after class.

Instructor is always available in class but only occasionally during office hours.

The instructor helps a student by allotting time after class for special problem sessions at a time convenient for the instructor as well as the student.

Instructor is available only in the classroom.

The instructor helps a student at a time that is convenient only to the instructor.

Instructor frequently breaks an appointment the student makes with the instructor.
### CLASS INTEREST - The ability of the instructor to keep the class's attention.

<table>
<thead>
<tr>
<th>Score</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>6</td>
<td>The instructor excites the student in class.</td>
</tr>
<tr>
<td>5</td>
<td>The instructor challenges the class so that few students are absent from class during the semester.</td>
</tr>
<tr>
<td>4</td>
<td>The instructor occasionally provides stimulating ideas in class.</td>
</tr>
<tr>
<td>3</td>
<td>The instructor has students in class who are frequently absent because the students find the class boring.</td>
</tr>
<tr>
<td>2</td>
<td>The instructor speaks with a monotone voice.</td>
</tr>
<tr>
<td>1</td>
<td>The instructor bores the class.</td>
</tr>
<tr>
<td>0</td>
<td>The instructor frequently stimulates thought and interest in the class.</td>
</tr>
</tbody>
</table>
RELATIONS WITH STUDENTS - the instructor communicates at an intellectual level that corresponds with the student.

- The instructor relates to the class so each student can follow what is going on at all times.
- The instructor explains new terms in a simple language for the student.
- The instructor speaks a vocabulary on the level of the student.
- The instructor’s lecture is at a pace that is too fast for the students to understand.
- The instructor is inconsistent in his level of approach with students.
- The instructor is unable to stay at the level of the student.
**INSTRUCTOR ATTENDANCE**—the instructor comes regularly to class and is on time

<table>
<thead>
<tr>
<th>Grade</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>7</td>
<td>The instructor is dependable. The instructor gives advanced notice for cancelled classes.</td>
</tr>
<tr>
<td>6</td>
<td>The instructor is in class when the student gets there and is the last one to leave.</td>
</tr>
<tr>
<td>5</td>
<td>The instructor is seldom absent, but does not announce his absences.</td>
</tr>
<tr>
<td>4</td>
<td>The instructor is frequently late for class.</td>
</tr>
<tr>
<td>3</td>
<td>The instructor frequently makes the class wait ten minutes before coming in.</td>
</tr>
<tr>
<td>2</td>
<td></td>
</tr>
<tr>
<td>1</td>
<td></td>
</tr>
<tr>
<td>0</td>
<td>The instructor is frequently absent and/or tardy.</td>
</tr>
</tbody>
</table>
CLASS PARTICIPATION - the instructor creates an atmosphere where the student can express his opinion.

- The instructor encourages discussion, questions, or student opinions
- The instructor explains questions the students may have.
- The instructor stops at various times during a lecture and asks for the questions.
- The instructor forms no idea on his own nor does he let the class.
- The instructor reads information from the text.
MOTIVATION OF INSTRUCTOR - the instructor shows interest in his subject.

The instructor is constantly researching and is up-to-date

The instructor brings up new points that relate with the material

The instructor admits when he does not know the answer to a student's question and provides the answer at the next lecture.

The instructor is interested in his subject but is not up-to-date

The instructor talks extensively about a subject but is unaware of new developments

The instructor does not know current material even though the instructor has taught the subject for a long time.

The instructor is bored with his own lecture and is outdated
SENSITIVITY TO CLASS PROGRESS- the ability of an instructor to follow a plan but be flexible.

<table>
<thead>
<tr>
<th>Score</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>The instructor provides no organization to the class and is insensitive to the student's needs.</td>
</tr>
<tr>
<td>1</td>
<td>The instructor states that he does not care if the students learn or not.</td>
</tr>
<tr>
<td>2</td>
<td>The instructor skips from one part of his lecture to another.</td>
</tr>
<tr>
<td>3</td>
<td>The instructor makes out a set plan. The plan covers the entire semester and the student knows what to expect.</td>
</tr>
<tr>
<td>4</td>
<td>The instructor follows a class outline as scheduled.</td>
</tr>
<tr>
<td>5</td>
<td>The instructor sticks to a class syllabus all semester. When the material was difficult, the instructor spent extra time on it.</td>
</tr>
<tr>
<td>6</td>
<td>The instructor follows a class outline that is structured but flexible to fit the student's needs.</td>
</tr>
</tbody>
</table>
**GRADING - the degree to which there is a meaningful policy in evaluation.**

<table>
<thead>
<tr>
<th>Score</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>7</td>
<td>The instructor gives extra credit for material covered outside of class.</td>
</tr>
<tr>
<td>6</td>
<td>The instructor knows the student's abilities. A good grade does not require 100% knowledge. The grade is determined by how the student has improved.</td>
</tr>
<tr>
<td>5</td>
<td>The instructor is occasionally inconsistent in his grading.</td>
</tr>
<tr>
<td>4</td>
<td>The instructor demands a great deal of work for a decent grade.</td>
</tr>
<tr>
<td>3</td>
<td>The instructor shows favoritism in giving grades rather than the knowledge the student acquires.</td>
</tr>
<tr>
<td>2</td>
<td>The instructor is unfair in grading.</td>
</tr>
<tr>
<td>1</td>
<td>The instructor gives tests only having one right answer that corresponds to the scoring key.</td>
</tr>
</tbody>
</table>
TEACHER PREPARATION - the instructor is ready to relate his worthwhile information in class.

<table>
<thead>
<tr>
<th>Score</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>7</td>
<td>The instructor covers the subject thoroughly</td>
</tr>
<tr>
<td>6</td>
<td></td>
</tr>
<tr>
<td>5</td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>The instructor covers the subject superficially</td>
</tr>
<tr>
<td>3</td>
<td>The instructor shows films to keep from lecturing</td>
</tr>
<tr>
<td>2</td>
<td>The instructor presents the facts but not the rationale behind the facts.</td>
</tr>
<tr>
<td>1</td>
<td></td>
</tr>
<tr>
<td>0</td>
<td>The instructor reads from the book and does not know what he is talking about.</td>
</tr>
</tbody>
</table>
Are there any areas of performance and/or unique behaviors which the instructor has contributed to the class which makes him effective or ineffective as an instructor?
References


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


