A Comparison of Illusory Halo Levels Between Conventionally-Developed and Factor Analytically-Developed Behaviorally Anchored Rating Scales

Robert Baker

Follow this and additional works at: http://digitalcommons.wku.edu/theses
Part of the Psychology Commons

Recommended Citation
http://digitalcommons.wku.edu/theses/1886
A COMPARISON OF ILLUSORY HALO LEVELS BETWEEN
CONVENTIONALLY-DEVELOPED AND FACTOR ANALYTICALLY-DEVELOPED
BEHAVIORALLY ANCHORED RATING SCALES

A Thesis
Presented to
The Faculty of the Department of Psychology
Western Kentucky University
Bowling Green, KY 42101

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

by
Robert Mitchell Baker
May, 1985
A COMPARISON OF ILLUSORY HALO LEVELS BETWEEN
CONVENTIONALLY-DEVELOPED AND FACTOR ANALYTICALLY-DEVELOPED
BEHAVIORALLY ANCHORED RATING SCALES

Recommended 6/6/55
(Date)

Ralph M. Mandel
Director of Thesis

John O’Connor

Elizabeth S. Effweyes

Approved 7-3-55
(Date)

Edgar Strong
Dean of the Graduate College
AUTHORIZATION FOR USE OF THESIS

Permission is hereby

☑ granted to the Western Kentucky University Library to
make, or allow to be made photocopies, microfilm or other
copies of this thesis for appropriate research or scholarly
purposes.

☐ reserved to the author for the making of any copies of this
thesis except for brief sections for research or scholarly
purposes.

Signed Robert Mitchell Baker

Date July 9, 1985

Please place an "X" in the appropriate box.

This form will be filed with the original of the thesis and will control
future use of the thesis.
TABLE OF CONTENTS

List of Tables ........................................ iv
List of Appendices ....................................... v
Abstract ..................................................... vi
Introduction ................................................ 1
Review of the Literature ................................... 5
Method ......................................................... 41
Results ......................................................... 48
Discussion ................................................... 53
Tables .......................................................... 65
References .................................................... 70
Appendices .................................................... 75
## LIST OF TABLES

<table>
<thead>
<tr>
<th>Table</th>
<th>Description</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>Table 1</td>
<td>Behavioral Incidents Composing Modified Bars</td>
<td>65</td>
</tr>
<tr>
<td>Table 2</td>
<td>Dimension Intercorrelations for the Harari-Zedeck BARS</td>
<td>66</td>
</tr>
<tr>
<td>Table 3</td>
<td>Dimension Intercorrelations for the Modified BARS</td>
<td>67</td>
</tr>
<tr>
<td>Table 4</td>
<td>Interrater reliabilities of the Harari-Zedeck BARS and Modified BARS by class</td>
<td>68</td>
</tr>
<tr>
<td>Table 5</td>
<td>Dimension Means and Standard Deviations Across Rates</td>
<td>69</td>
</tr>
</tbody>
</table>
## LIST OF APPENDICES

<table>
<thead>
<tr>
<th>Appendix</th>
<th>Description</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>Appendix A</td>
<td>Harari-Zedeck BARS</td>
<td>75</td>
</tr>
<tr>
<td>Appendix B</td>
<td>Modified BARS</td>
<td>85</td>
</tr>
</tbody>
</table>
A COMPARISON OF ILLUSORY HALO LEVELS BETWEEN
CONVENTIONALLY-DEVELOPED AND FACTOR ANALYTICALLY-DEVELOPED
BEHAVIORALLY ANCHORED RATING SCALES

Robert Mitchell Baker

May 1985

Directed by: Ray Mendel, Betsy Erffmeyer, and John O'Connor
Department of Psychology
Western Kentucky University

The purpose of this study was to compare halo levels between a set of behaviorally anchored rating scales (BARS) developed by conventional methodology with a set of BARS developed by factor analytic means. The conventionally-developed BARS were developed at the University of California-Berkeley by Department of Psychology faculty members in conjunction with undergraduate psychology students. The factor analytically-derived BARS were developed during the course of the research reported here. In Phase 1 of the present study, undergraduate psychology students rated their instructors on a 1-7 Likert scale using the individual anchors which comprised the conventional BARS. A factor analysis of these ratings was used to form a modified set of BARS. In Phase 2, a new group of undergraduate psychology students was asked to rate their instructors using either the conventional (Harari-Zedeck) BARS or the newly developed modified BARS in order to test for differences in the halo levels for each form of BARS. Halo was measured
by the strength of the dimension intercorrelations in each set of scales. It was hypothesized that the factor analytic approach would yield a set of scales with reduced levels of illusory halo. Results indicated, however, that the conventionally-developed and factor analytically-developed BARS were not significantly different in their halo levels. A post hoc reliability check revealed that the conventionally-developed BARS had significantly less interrater reliability than the factor analytically-derived BARS, which may partially account for the lack of difference in the scales' halo levels.
INTRODUCTION

The development of adequate measures of job performance has traditionally been a primary concern of industrial psychologists. Since many such measures of performance take the form of subjective ratings (often by a supervisor or by peers), a great deal of time and effort have been devoted to the development of measures which seek to minimize human error in rating performance. Most of this research has been directed toward the identification and minimization of traditional rater errors such as leniency, central tendency and halo error (Landy, Vance, Barnes-Farrell, & Steele, 1980).

Halo error can be defined as a global evaluation biasing ratings of individual dimensions of performance, resulting in inflated interdimension correlations (Burnaska & Hollman, 1974). Despite the amount of attention this problem has received, it has remained a particularly stubborn one. Recent overviews of halo research still document the ubiquitous nature of halo (see Cooper, 1981a) more than six decades after Thorndike (1920) reported that the intercategory correlations resulting from supervisors' multidimensional ratings of subordinates' performance were "all higher than reality" (p. 25). Efforts directed toward rater training and scaling refinements have failed to alleviate the problem (Cooper, 1981b).
Behaviorally anchored rating scales (BARS) based upon scaled expectations (Smith & Kendall, 1963) have been touted as having good potential in overcoming many of the traditional rater errors, including halo (see Dunnette, 1966). BARS have been designed to reflect the major relevant dimensions of job performance, using specific behavioral incidents as anchor points along rating scales which represent these dimensions (Arvey & Hoyle, 1974). These behavioral incidents are arranged along a rating continuum so that the behaviors near the bottom of the continuum are reflective of average performance, and those near the top are reflective of highly effective job performance. By using concrete, commonly observable behaviors as anchors or reference points along the scale continuum, different raters "can be reasonably expected to share some common core of experience and values concerning behavior on the jobs they will rate" (Smith & Kendall, 1963, p. 150). Since ratings from different raters are usually combined in the typical performance evaluation setting (i.e. ratings from different raters are treated as though they were equivalent), it is hoped that this procedure provides all potential raters with a common frame of reference.

Although a more detailed version of BARS developmental procedures will be addressed in the review of the literature, the reader should aware that Smith and Kendall felt that in order to encourage raters to complete their ratings honestly and carefully, their participation in the scales'
development was essential. Group consensus regarding the number of performance dimensions, their definitions and the specific behavioral incidents which compose them should enable the raters to clearly see the distinction between the different dimensions, thereby helping them to avoid halo.

There is some indication, however, that these "intuited" dimensions developed in the BARS process may also be susceptible to halo error in the actual rating situation (Dickinson & Tice, 1977). One possible explanation may rest with BARS methodology itself. Any of the behavioral incidents derived by the potential raters may tap a performance dimension (or dimensions) beyond the one that item was intended to illustrate (Dickinson & Tice, 1977). It may be that behavioral incidents illustrative of more than a single dimension may cause the separate performance dimensions to seem less distinct, thereby resulting in spuriously high interdimension correlations, or halo.

In the present research, ratings on the behavioral incidents which composed a BARS designed to measure the effectiveness of a university psychology instructor (Harari & Zedeck, 1973) were subjected to a factor analysis. The resulting information was used to form a modified version of BARS, one containing fewer behavioral incidents redistributed into factors of job performance. Intercorrelations among scale ratings derived from this modified BARS will be compared to the intercorrelations among ratings obtained from the original BARS. The distinctiveness among the rated dimensions
is measured by their intercorrelations. Thus, if the intercorrelations among the factors of the modified BARS are lower than those among the intuited dimensions of the original BARS, halo error will be decreased.
REVIEW OF THE LITERATURE

As previously stated, research into the phenomenon labelled "halo error" is plentiful. In the following pages, an attempt will be made to define this error, describe how it has been measured, and to give the reader some notion of how and why BARS are seen as a useful tool in dealing with halo.

Halo error

According to Smith (1976), halo represents the failure of the rater to differentiate among ostensibly different aspects of job performance. This failure results in a rating on one aspect of performance "spilling over to affect ratings on another, resulting in high intercorrelations among ratings" for supposedly different dimensions (p. 757). Thus, halo may be thought of as the result of a global evaluation biasing the ratings on each performance dimension in the same direction, the result of which is inflated scale intercorrelations (Burnaska & Hollman, 1974). Halo may bias ratings in either a positive or negative direction. It merely indicates a lack of discrimination on the part of the rater (Smith, 1976).

In accordance with the amount of research into the topic of halo error, many methods of detecting this phenomenon
are currently in use. Cooper (1981a) has identified five such methods. Probably the most common is Thorndike's (1920) method of calculating intercategory correlations and drawing inferences about whether they are too high compared with some measure of "true" intercategory correlations. The problem of determining the magnitude of actual interdimension covariance, sometimes referred to as "true" or "absolute" halo (Cooper, 1981a), will be addressed later in this paper.

A second method estimates halo by examining intraratee variance across dimensions, and again inferences are drawn regarding whether the variances are too low compared to some model of true variance. A third method identifies the intercategory factor structure, detecting halo in cases where the factor structure is dominated by a general factor accounting for a sizable portion of the variance. A fourth method employs an analysis of variance design to detect halo. After submitting the data to a rater x ratee x dimension ANOVA, halo is said to be present if a rater x ratee effect is found. A fifth method examines the effects of manipulated categories on non-manipulated category ratings. Inferences about the presence of halo are then made from the results of this experimental method.

Sources of halo

Although a great deal of time and effort have been devoted to identifying and controlling halo, less is known about its possible sources. Cooper (1981a) lists six sources which are implicit in the methods used to detect halo. These
sources are 1) undersampling, 2) engulfing, 3) insufficient concreteness, 4) insufficient rater motivation or knowledge, 5) cognitive distortions, and 6) correlated true scores.

It should be noted that this last source is "true halo." The first five are sources of halo error, also referred to as "illusory halo" (Cooper, 1981a).

Undersampling as a source of halo is the result of raters basing his/her judgements on an insufficient number of observations. It is assumed that since the rater has not had sufficient opportunity to observe behaviors relevant to all rating dimensions, he/she must rely somewhat on a global impression which results in halo error.

Similarly, halo error resulting from engulfing is due to the rater's perception of overall job effectiveness influencing the individual dimension ratings. To the degree which raters believe that dimensions co-vary with a general impression of job effectiveness, dimension ratings are engulfed. This differs from undersampling in that halo error may persist even with increased opportunity to observe ratee behavior because of the strength of the raters' belief that the rated dimensions co-vary with their general impressions.

Rating dimensions which are insufficiently concrete may cause the rater to combine vaguely related observations, thus producing halo. Both dimension descriptions and the behavioral incidents which anchor the scales should be "empirically derived and sufficiently descriptive and concrete, as opposed to abstract" (Cooper, 1981a, p. 220). Scales
lacking these qualities tend to be partially redundant and overlapping.

The fourth source of halo error cited by Cooper stems from either carelessness or a simple lack of effort on the part of the rater. It is possible that the rater simply wishes to complete the ratings as quickly as possible, giving all ratees the same (or a highly similar) rating for all dimensions, which results in heavily haloed ratings.

The cognitive distortions perspective attributes halo error to faulty information storage and recall by the rater. According to this perspective, ratings are haloed because observations of past behaviors are not recalled with complete accuracy. What cannot be recalled from memory is replaced by the rater's belief about the dimensions' true covariances. What is recalled now are observations of ratee behavior distorted by the rater's perceptions of the degree to which the dimensions co-vary. If these perceptions are overestimates of true interdimension covariance (i.e., if the dimensions are not as highly intercorrelated as the rater believes them to be), the result will be halo error.

The final source of halo to be discussed here presents an entirely different kind of problem for the researcher. As stated above, true halo does not represent a form of rater error, but instead reflects the degree to which job traits or abilities are correlated in the real world. Since many jobs represent "a relatively homogeneous cluster of work tasks" (Dunnette, 1966), it is reasonable to assume
that the abilities are what the rating dimensions are designed to tap. Thus, it is not surprising that these dimensions will be correlated to some degree. The researchers' problem is determining the levels of true halo.

Due to the fact that research into ability taxonomies is still in the exploratory stage, the answer to the above question is still unknown (Dunnette, 1976). The use of "true scores" derived from ratings by "expert raters" may provide some help (see Borman, 1977, 1979), but the use of expert raters is not feasible in many research settings. Hopefully, the future development of feasible methods of determining true interdimension covariance will provide future studies with a solution to this problem.

As can be gathered from Cooper's review, halo error can originate from numerous sources. However, a common theme links several of these sources. With the exception of undersampling and insufficient rater motivation, other sources of illusory halo stem from the inability of the rater to conceptually distinguish one dimension from another. This inability does not imply that the rater cannot understand the definition of each dimension; rather he/she tends to see each dimension as part of the same ability or trait and thus relies on a more general, overall impression on which to base his/her ratings.

The development of BARS was intended to provide a method of overcoming these problems. One of BARS' purposes was to produce conceptually independent dimensions, which partly
through their development procedure elicit agreement among
the raters regarding both the distinctiveness of the separate
dimensions and the exhaustiveness of the domain of behaviors
exhibited on the job (Harari & Zedeck, 1973). The following
pages will outline BARS developmental steps and describe
how they were designed to overcome some of the sources of
halo error described above.

**BARS Developmental Steps and Halo Reduction**

While there have been minor modifications of the origi-
inal BARS methodology (Smith & Kendall, 1963), develop-
typically includes five steps. They are 1) generation of
behavioral incidents, 2) clustering of behavioral incidents
into performance dimensions, 3) retranslation of these inci-
dents into their original dimensions by a separate group
of participants, 4) scaling the incidents to represent the
degree of performance effectiveness each one typifies, and
5) the inclusion of items on the final scales (Schwab,
Heneman, & DeCotiis, 1975). This format was chosen in order
to combine the advantages of direct observation of ratee
behavior with the acceptability to raters of graphic rating
scales (Smith & Kendall, 1963).

**Generation of behavioral incidents.** In the initial
stage of BARS development, incidents which exemplify poor,
average and good performance on the job are generated by
persons familiar with that job, typically those who will
be using the scales themselves. The advantage here is that
BARS uses input from individuals who know the job and its
requirements thoroughly. The goal is to achieve group consensus regarding the exhaustiveness of the domain of job behaviors, which serves to increase the content validity of the scales (Schwab et al., 1975). Ideally, this process of reaching group consensus results in improving both the face validity of the scales and the domain of the performance areas to be evaluated, which in turn should result in conceptually distinct and independent dimensions meaningful to the rater (Harari & Zedeck, 1973). Theoretically, the distinctiveness (and hence the independence) of these dimensions should aid the rater in avoiding halo.

Smith & Kendall believed that potential rater participation in scale development was essential for two reasons. First, raters had to be "sold" upon the desirability of completing the ratings honestly and carefully. They believed that the apparent usefulness of the procedure (the face validity of BARS) would encourage raters to take their time and complete the ratings to the best of their abilities. In fact, Smith and Kendall cite the unacceptability to the rater of traditional rating formats as the most crucial deciding factor in the decision to develop a new rating format. Secondly, it was believed that rater participation was critical in obtaining scales with sound psychometric properties—specifically, higher intrarater reliability and independent dimensions.

Two recent studies (Dickinson & Zellinger, 1980; Ivancevich, 1980) seem to indicate that raters prefer BARS
over more traditional rating formats. Dickinson and Zellinger found that students rating faculty teaching performance felt that BARS were more successful than a mixed-standard format in meeting assessment goals and providing feedback to students and faculty members. Also a significantly greater percentage of students reported that they preferred to use BARS over the mixed-standard format. Ivancevich (1980) found that engineers rated by BARS reacted more favorably to performance evaluations than those who were rated by a set of trait scales the company already had in use. Engineers rated by BARS also reported more meaningful feedback, more clarity, and less job related tension than engineers using trait scales.

A more recent study (Silverman & Wexley, 1984) revealed that hospital employees who had the opportunity to participate in the development of BARS perceived their subsequent performance appraisal interviews differently than those not participating in scale development. Ratees who participated in BARS development perceived the interview to be more useful, that goals and objectives were set to a greater extent, and that they contributed more to the interview. They also reported being better satisfied with the interview and were more motivated to improve their performance. Unfortunately, no measures were available to determine if there were actual improvements in subsequent job performance. Apparently then, BARS are seen as highly useful scales, which should encourage raters to use them honestly and carefully.
With regard to the belief that rater participation in scale development results in psychometrically superior scales, the results from experiments are less encouraging. Keaveny and McGann (1975) found no difference between participants and non-participants in scale construction with respect to leniency, precision of ratings, and halo error. In a similar study, Bernardin, LaShells, Smith and Alvares (1976) found no differences in halo levels for a BARS which had been developed by the eventual raters and one developed by the experimenters. Both of these studies used student ratings of faculty teaching performance as data sources. In another study concerning the effectiveness of ROTC instructors in the classroom (Friedman & Cornelius, 1976), a significantly smaller amount of illusory halo was present in the ratings of those raters in the participatory condition. Based upon these and other contradictory findings, a review of the BARS literature by Jacobs, Kafry and Zedeck (1980) points out that "the merits of participating in scale construction are equivocal and any conclusions regarding participation effects would be premature." (p. 630).

These results seem to indicate that although raters may prefer to use BARS over other rating scales and see BARS as having more face validity, the preference apparently is not enough to ensure that ratings will show less halo with rater participation. However, to the degree that halo is a result of rater carelessness or lack of effort, BARS may very likely encourage the careful and honest use of rating scales, thus reducing halo.
Development of performance dimensions. In the second stage of scale development, the behavioral incidents are clustered into a smaller set of performance dimensions and given formal definitions. This step is usually carried out by the same group of participants who took part in the first step. The identification and definition of performance dimensions by job-knowledge participants was originally the first step of BARS development, but nearly all studies since then use this procedure as the second step in order to keep the participants focused on specific rather than global examples of job behavior (Schwab et al., 1975).

Having potential raters participate in the clustering of behavioral incidents into performance dimensions plays an important role in helping raters to distinguish one dimension from another. BARS developmental procedure "gambles that among a fairly homogenous group of judges" (whether they be managers, head nurses, or students in a psychology class) beliefs about what constitutes poor, average and above average job behaviors will be reasonably well-standardized (Smith & Kendall, 1963, p. 151). These beliefs, operationalized in the form of written behavioral incidents, are "organized into areas not by theoretical similarity, but by judged similarity as indicated by the raters, and that the areas represent dimensions meaningful to the raters" (p. 151). Thus, the process of generating incidents and clustering them into performance dimensions should aid the rater in conceptualizing these dimensions as separate entities, thereby reducing halo.
Research addressing this issue has yielded mixed results. Borman and Dunnette (1975) found that BARS with behavioral anchors yielded ratings with less halo than similarly developed BARS without anchors. However, when these two formats were combined and compared to a set of trait scales, this difference disappeared. Since both of the BARS conditions contained identical dimension definitions, it cannot be determined if the process of having raters cluster and define the incidents will reduce halo.

In a study examining the use of dimension clarification statements, Bernardin, LaShells, Smith and Alvares (1976) found no difference between a BARS with and a BARS without clarification statements with regard to halo. Despite these results, Bernardin et al. recommended that potential raters develop dimension clarification statements in order to inhibit leniency error and increase rating discriminability across raters. Rating discriminability was defined as a significant difference in mean standard deviations of ratings across all performance dimensions.

In a study comparing BARS to a mixed-standard scale format (Finley, Osburn, Dubin, & Jeanneret, 1977), results showed BARS composed of both specific and nonspecific behavioral incidents yielded significantly less halo. Unfortunately, the authors do not state whether the raters participated in BARS development; thus it is not possible to determine if rater development of behavioral incidents and performance dimensions were responsible for lowered amounts of halo.
The goal of this second stage of BARS development is to help the potential rater to see the dimensions as concrete, observable entities. Dimension titles which may seem vague and highly subjective to the rater are given concreteness and distinctiveness through their definitions and behavioral incidents. To the degree that raters tend to give haloed ratings in situations where the dimensions are vague and undefined, BARS aid the rater in avoiding halo by giving the rater a clearer understanding of what he/she is being asked to rate.

Retranslation. In the third step of BARS development, a separate group of potential raters are asked to retranslate (or reallocate) the behavioral incidents into the performance dimensions they believe each incident best represents. In order to qualify for inclusion on the final set of scales, an item must pass a retranslation criterion. Typically, an item is retained if a certain percentage (usually 50-80 percent) of the raters assign an incident to the same dimension (Schwab et al., 1975). Items which pass the criterion are said to have been successfully retranslated. This process is similar to the one undertaken when material is translated from one language to another. Material is translated into a foreign language and then retranslated by a different interpreter back into the original language, thus assuring that the translation adheres faithfully to the original (Campbell, Dunnette, Lawler & Weick, 1970). In terms of BARS development, retranslation assures that
the meanings of both the performance dimensions and behavioral incidents are "highly specific and non-ambiguous" (p. 120).

Dickinson and Tice (1973) have identified the key elements of the retranslation step. They are 1) achieving face validity in the scales through participant involvement, 2) the existence of a group of potential raters whose views of the job and its required behaviors are relatively homogeneous, and 3) the use of an "intuitive factor analysis" to sort these specific behavioral incidents into job dimensions.

The reader may notice that these first two elements can be considered true of the BARS developmental process as a whole. The input from a relatively homogenous group of potential raters is elicited in the retranslation phase as well as in the preceding two. The last element, the intuitive factor analysis, is involved in both the development of performance dimensions and the retranslation phases, where items are clustered into dimensions and then rejected or confirmed. The point here is that BARS development is an iterative process in which the potential rater is thoroughly familiarized with the scale. The fact that the rater was not simply a passive observer but instead an active participant in scale construction will help "sell" the rater on completing the scales accurately and honestly (Smith & Kendall, 1963). In combination, these first three steps should yield conceptually distinct performance dimensions, each of which represents a factor of job performance with
a set of behavioral examples to define and measure that factor (Dickinson & Tice, 1977). In the actual rating situation on the job, the procedures should translate into lower inter-dimension correlations, an indication of reduced halo. Field research conducted to test this premise will be discussed later in this paper.

Scaling incidents. The same group of potential raters who participated in the retranslation step are usually involved in this step. They are asked to rate each behavioral incident (typically on a 7- or 9-point scale) as to the level of effectiveness that particular incident represents. The average rating the group assigns to an incident is taken as its level of effectiveness. The standard deviation of the item's ratings serves as a criterion for that item's inclusion on the final scales. Typically, items that have standard deviations of less than 1.5 are retained (Schwab et al., 1975). Chosen are items which cover the entire range of the scale continuum, and the item with the lower standard deviation is chosen for placement on the final scale when two or more have the same scale value.

This step of the BARS developmental process serves two purposes. First, the process of having potential raters scale each behavioral incident is sound psychometric practice. Assuming that a relatively homogenous group of potential raters are involved, the process of scaling incidents should yield reliable estimates of each incident's scale value, assuming a sufficient number of potential raters are sampled.
It makes sense that a carefully constructed performance measure should not only result in lower amounts of halo, but reduced amounts of other types of rater errors as well (Smith & Kendall, 1963). Secondly, this phase of BARS development allows yet another opportunity for potential raters to become involved in the scales' construction. The potential raters must be in agreement on the dimensions which cover the domain of job performance and the behavioral incidents which illustrate these dimensions. Without agreement regarding the degree of effectiveness each incident represents, the scales are of little or no value.

As outlined by Smith and Kendall, raters should be part of a relatively homogenous group; in practice, however, they may come from different organizational levels, have different levels of job experience, and differ in their amount and area of education. Potential raters involved in the process are not identical in their perspective of behaviors constituting effective job performance, and this difference in perspective has implications for the scaling phase of BARS development. To the degree groups of potential raters have meaningful differences in the value they attach to a particular job behavior, the reliability of the scales is weakened.

In a study comparing ratings of nursing performance compiled by both supervisors and head nurses (different organizational levels), Zedeck and Baker (1972) found significant agreement (convergent validity) on all five rated dimensions. Agreement was far from perfect, however, and this
discrepancy was explained by the fact that the two groups had differential opportunities to observe ratee behavior. It was suggested that since the head nurse group contributed nearly twice as many behavioral incidents as did the supervisors, head nurses were more familiar with ratees' behaviors and were in a better position than the supervisors to make accurate ratings. Campbell, Dunnette, Arvey and Hellervik (1973) found greater agreement among ratings made by store managers when compared with ratings made by assistant managers. The authors believed that the low interrater agreement was due to the fact that the job of assistant manager was geared more toward merchandising than supervision, thus depriving assistant managers from having ample opportunity to observe ratee behavior. Zedeck, Imparato, Krausz, and Oleno (1974) conducted a study in which the scale values of behavioral incidents as judged by a supervisory group (head nurses) were compared to the scale values of the same incidents as judged by a subordinate group (registered nurses). Zedeck et al. hypothesized that the subordinate group would rate the same set of incidents as indicative of better performance than would the supervisory group, and the results of the experiment confirmed this hypothesis. The authors suggested that when such a situation arises the two groups should meet and discuss their differences regarding their perceptions of effective job behaviors. Blood (1974) suggested that such a situation might indicate poor communication between organizational levels, and differences between
those two groups should be discussed and resolved in order to achieve smooth organizational functioning. In a similar vein, Beatty, Schneier, and Beatty (1977) found that ratees (computer analysts) perceived undesirable job behaviors as occurring less often and desired behaviors more often than did raters (supervisors). Both of these groups agreed on mid-scale items (moderately effective job behaviors). The authors suggested that both raters and ratees may be biased in their recall of behaviors due to traditional adversarial roles between management and labor.

Cascio and Valenzi (1977) found that raters with more job experience gave higher mean ratings than raters with less experience. Raters with less education tended to give higher mean ratings than did more educated raters. However, none of these differences were practically significant, thus these results must be considered tentative.

The results of these studies seem to suggest that perceptions of job effectiveness are not completely standardized among groups of potential raters, although remedies for this problem seem to be within reach. Two of these studies (Campbell et al., 1973; Zedeck & Baker, 1972) showed results that rater differences were probably the result of differential opportunity to observe behavior. The obvious solution here would be to ensure that all potential raters have ample opportunity to observe ratee behavior before the scales are administered. Two studies (Beatty et al., 1977; Zedeck et al., 1974) revealed differences between raters at different
organizational levels. Rater differences in the assigned values of anchors as moderated by experience and education were not large enough to be practically significant (Cascio & Valenzi, 1977).

In summary, it seems that although using potential raters to scale behavioral incidents does not result in perfect agreement regarding scale values, the benefits of this procedure more than offset the drawbacks. Potential raters are involved in another important part of scale development, and thus may be more motivated to complete the ratings accurately. This process may provide the additional benefit of pointing out to an organization where differences of opinion may exist regarding the effectiveness of certain job behaviors. This identification of points of disagreement is one of the many beneficial by-products of the BARS developmental process (see Blood, 1974).

Final instrument. The behavioral incidents which meet the criteria set forth in the retranslation and scaling steps are used as anchors for the performance dimensions. Typically, six or seven incidents anchor each dimension, and there are usually six to ten dimensions in a BARS. Each dimension is represented by a vertical scale with the retained behavioral incidents located along the continuum depending upon their value established in the preceding step (Schwab et al., 1975).

This step completes the BARS developmental process. Actually, this final phase is nothing more than the culmination
of the first four, but it allows those who participate in its construction to see the product of their labor. As stated above, Dunnette (1966) believes that participation may encourage raters to increase their commitment to making accurate assessments. According to Smith and Kendall (1963), this commitment is the essence of BARS.

BARS has the additional advantage of yielding an "intuitive factor analysis" which clusters behavioral incidents into performance dimensions. The definitions of these dimensions and the behavioral incidents which illustrate them should enable the rater to identify the dimension he/she is rating in concrete, observable behaviors. To the extent that halo is a result of insufficient concreteness of rating dimensions, BARS again should be useful in reducing levels of illusory halo.

Another variable affecting the level of illusory halo is the reliability of the scales. It is possible to reduce halo levels simply by creating a less reliable set of scales. In a situation where a set of scales has little or no reliability, correlations among dimensions will be virtually zero. Thus, it may appear that halo among performance dimensions has been reduced when in reality the reduction is due to the lack of scale reliability. Due to the fact that past research has used a large variety of reliability measures (see Jacobs, Kafry & Zedeck, 1980, pp. 627-629) and a similar variety of halo measures, it is difficult to tell whether reliability differences in past BARS research is responsible for inconsistent halo levels.
The preceding pages have addressed the issue of how the individual steps of BARS development can help to reduce halo levels, and research investigating BARS' ability to lower halo through rater participation has been cited. Discussed in the next section will be research comparing BARS to other types of rating formats in order to determine if BARS is a better alternative in the search for a method of reducing halo.

**BARS versus alternative methods of rating performance**

Although a vast number of studies have examined BARS methodology and halo reduction, much of this research has involved comparing one variation of BARS methodology with another (rater participation versus non-participation in scale construction, inclusion versus omission of behavioral anchors, etc.). Many of these studies were described above. This section will discuss the performance of BARS versus alternative rating formats such as mixed-standard and summated rating scales.

The majority of research comparing BARS to alternative formats has compared BARS with some type of graphic rating scale usually composed from dimension definitions elicited during BARS development. Only three studies (Borman & Dunnette, 1975; Burnaska & Hollman, 1974; Friedman & Cornelius, 1976) have compared BARS with totally independent rating formats. Burnaska and Hollman (1974) compared BARS to a set of summated rating scales designed to measure teacher effectiveness. A slight reduction in halo levels was found
in the BARS condition, but this effect accounted for only about 3 percent of the variance and therefore cannot be considered meaningful. There was no evidence that summated scales composed of the same dimensions as BARS but using a Likert-type rating format were any more susceptible to halo than a traditional BARS.

In another study comparing two types of BARS to a series of scales involving trait-oriented dimensions used by the U.S. Navy to rate the performance of their officers (Borman & Dunnette, 1975), the trait scales showed more halo than either a traditional BARS or a BARS without the behavioral anchors included. Once again, the differences were not practically significant, accounting for only 5 percent of the variance. There were no differences among the two types of BARS with regard to halo.

Although the main thrust of the research conducted by Friedman and Cornelius (1976) was to check the effects of rater participation in scale construction on halo levels, a BARS was compared to a graphic rating scale developed by potential raters. This graphic scale used a Likert-type format to measure the performance of ROTC instructors. Only performance dimensions were developed for the graphic behavioral incidents, but merely rated subjects on each dimension using a 1-7 scale. Halo accounted for only 1 percent of the total variance in the BARS condition and only 3 percent in the graphic rating scale condition. This difference was neither statistically nor practically significant.
In numerous studies, BARS has been compared with summed rating scales based upon data generated by BARS methodology (Arvey & Hoyle, 1974; Bernardin, 1977; Berardin, Alvares, & Cranny, 1976; Borman & Vallon, 1974; Campbell, Dunnette, Arvey & Hellervik, 1973). Campbell et al. (1973) sought to determine if behaviorally-based rating scales would contain lesser amounts of halo than a rating procedure that was not behaviorally anchored. Using data gathered from the performance evaluations of department managers of a large retail store chain, it was found that behaviorally anchored scales yielded ratings with less leniency error and less halo. Items on the summed scale came from favorable behavioral incidents (those above the scale midpoint) which were then rated on a Likert-scale format. Both sets of scales shared identical dimension titles and definitions. Likewise, Borman and Vallon (1974) compared BARS to a similar format which lacked only the behavioral anchors in order to check the effectiveness of BARS in a setting different from the one in which it was developed. Halo levels were found to be virtually equivalent for the two formats. The authors suggested that BARS' failure to demonstrate superiority may have been the result of out-of-date behavioral examples. Additionally, raters did not have the opportunity to prepare their own dimensions and anchors, which may have reduced any effects in the BARS condition.

Bernardin, Alvares, and Cranny (1976) compared BARS to two types of summed rating scales: one composed from
the major components of the definitions of the performance
dimensions and another containing the same incidents sur-
viving an item analysis (the former summed scale, SRS 1,
contained 39 behavioral incidents; the latter, SRS 2, con-
tained 24 incidents). Halo levels were significantly lower
for BARS than for SRS 1, but this difference was reduced
to a nonsignificant one when BARS was compared to SRS 2.
The authors suggested that more rigorously developed scales
(i.e., those scales subjected to an item analysis) should
result in scales possessing superior psychometric properties.
In a similar study, Bernardin (1977) compared ratings from
BARS to two different summed rating scales. The first
set of summed scales were comprised of components of the
definitions of performance dimensions rated on a 1-7 scale
with anchors ranging from "always" to "never." The second
set of summed scales contained behavioral incidents which
had survived all phases of BARS development. There were
no significant differences in leniency, interrater reliab-
ility or halo for any of the three formats. Results of this
study revealed lower overall levels of halo relative to pre-
vious studies, but this finding may not be surprising in
light of the fact that all three formats used "optimal" BARS
development procedures (see Bernardin, LaShells, Smith &
Alvares, 1976).

BARS have also been compared to mixed-standard formats
in three other studies (Arvey & Hoyle, 1974; Finley, Osburn,
(1974) compared a BARS to a mixed-standard scale composed of behavioral incidents surviving all phases of BARS development but not included on the final scales. For the mixed-standard scale, behavioral incidents were presented in random order and raters were asked to judge if the ratee's behavior was better than, the same as, or worse than each behavior listed. There was no significant difference between the two methods with regard to halo.

Finley et al. (1977) compared two different types of BARS (one with behaviorally-general anchors and one with behaviorally-specific anchors) to a mixed-standard scale utilizing the same behavioral descriptions contained in the behaviorally-general scales, but mixed in random order for presentation to raters. Results from this study were inconclusive; the BARS with specific anchors yielded ratings with slightly less halo than did the mixed-standard format, but the BARS using general anchors fared slightly worse than the mixed-standard scale. None of these differences were statistically significant.

Saal and Landy (1977) obtained performance ratings of police officers in seven different cities. Ratings were made using both BARS and mixed-standard scales derived from the behavioral incidents anchoring the BARS. Surprisingly, results showed that BARS' ratings contained significantly larger amounts of halo error than did the mixed-standard scale, although the BARS was found to have greater inter-rater reliability.
According to their review of BARS research, Kingstrom and Bass (1981) found no clear superiority for BARS over alternative rating methods with regard to halo. Of the more than 50 comparisons between BARS and alternative formats (many studies included several comparisons each), about 70 percent (35 of 51) found BARS to be less susceptible to halo, while about 28 percent (14 of 51) found BARS to be more susceptible. While these percentages may seem to suggest that BARS may have excellent potential for reducing illusory halo, these differences were generally small ones, seldom being statistically or practically significant (Kingstrom & Bass, 1981). Thus while research generally suggests that BARS are somewhat less susceptible to halo error, conclusions regarding their absolute superiority to alternative methods are tentative at best.

Explanations for the ubiquitous nature of halo

As mentioned above, a certain amount of true halo will always exist among ostensibly independent dimensions of performance. The value of the intercorrelation between any two measures of cognitive ability is not zero but some small positive value (Hulin, 1982). In a job setting where most of the tasks are of a homogenous nature and thus basically require one general type of skill or ability, performance dimensions are likely to be highly correlated only because each rated dimension is likely to be an outgrowth of the same skill (Cooper, 1981a). This phenomenon is known as "true halo" and describes the degree of dimension covariance
as it exists in the real world. Although the measurement of true halo poses a difficult problem for researchers, it cannot be ignored and efforts to statistically eliminate halo through partial correlation techniques (see Landy, Vance, Barnes-Farrell & Steele, 1980) have met with the criticism that such techniques partial out true dimension covariance (true halo) as well as illusory halo (Harvey, 1982; Hulin, 1982; Murphy, 1982). Although some methods of estimating true halo are in various stages of development (e.g., the use of expert raters or data from multitrait-multimethod matrices), true halo measures are difficult to obtain. Thus, most current researchers must be satisfied with comparing various forms of rating scales and concluding that those with lesser degrees of halo are superior.

The vast majority of halo research, however, has been focused on illusory halo. This research suggests that the great promise BARS once held as a method of halo reduction has never been fully realized (Kingstrom & Bass, 1981). Clearly then, additional sources of halo must still be operating, sources unaffected by BARS developmental procedures. The reader may remember that among Cooper's five sources of illusory halo (undersampling, engulfing, insufficient concreteness, carelessness/lack of effort, and cognitive distortions) only two of these (insufficient concreteness and carelessness/lack of effort) are sources which are directly confronted in BARS development. Thus, undersampling, engulfing and cognitive distortions remain as possible sources
of halo in BARS use -- a possible explanation of why BARS has shown only very slight improvements in halo reduction over alternative methods.

Of these three remaining potential sources of halo, undersampling seems to be least likely affected by the form of rating scale used to evaluate job performance. Halo due to undersampling results from the rater's having insufficient opportunity to observe ratee behavior. When the rater has not had the opportunity to observe a sufficient amount of ratee behavior, he/she may be forced to base ratings on nothing more than a general impression, which results in haloed ratings. Since research seems to suggest that raters having less contact with ratees showed greater amounts of halo in their ratings (Campbell et al., 1973; Zedeck & Baker, 1972), it makes sense that only someone with sufficient opportunity to observe the ratee's job behavior should be asked to rate. This advice is not peculiar to BARS, but is a fundamental principle which should be considered regardless of the type of scale being used.

The two remaining sources of halo, engulfing and cognitive distortions, are similar in that both sources trace halo back to the rater's preconceived notions of dimension covariance (Cooper, 1981a). According to Cooper, engulfing occurs to the extent which raters believe that dimensions co-vary with general impressions of overall job effectiveness. Similarly, the cognitive distortion perspective suggests that when details of ratee behavior are forgotten by
the rater, these details are replaced with the rater's beliefs regarding dimension covariances. To the degree that the rater overestimates true dimension covariance, the ratings will be haloed (Cooper, 1981a).

Inherent in the BARS approach is the assumption that raters can accurately discriminate among a large set of stimuli. The typical BARS contains 6-10 performance dimensions and approximately seven behavioral anchors per dimension. The rater is asked to read each behavioral anchor and compare the ratee's typical job behaviors to these anchors. Thus, a rater may be asked to make as many as 70 discriminations (10 dimensions x 7 anchors) for each ratee. Although it may be true that BARS aid the rater by defining the performance dimensions in terms of concrete, observable behaviors, they may burden the rater by requiring many specific judgements and discriminations in the perception of job behaviors (Schneier, 1977). Perhaps the cognitive ability of the rater may affect his ability to make the necessary discriminations required by BARS, which in turn affects the amount of halo present in the ratings. It has been suggested that an individual's ability to process information "might be related to the number of evaluation dimensions used for performance assessment. It may be that the greater the number of dimensions, the more difficult the discriminability" (Kafry, Jacobs, & Zedeck, 1979, p. 188). In Smith and Kendall's original article, it was stated that "there are too many scales for easy handling and the number of qualities
(dimensions) can be reduced after actual field use permits computation of scale intercorrelations" (Smith & Kendall, 1963, p. 154).

Dickinson and Tice (1977) offer a slightly different explanation as to why dimensions are seen by raters as more highly correlated than they are in reality. Rather than blaming halo on an excessive number of required discriminations, they believe that many of the behavioral examples are illustrative of more than one performance dimension, thereby blurring the distinction between dimensions. The reader will remember that during the retranslation phase of BARS development, only 50-80 percent agreement by potential raters is required to successfully retranslate a behavioral item to its original dimension. Therefore, even after meeting several criteria in the various stages of scale development, any given behavioral item may reflect more than one dimension in the eyes of some raters. Such an item may be "factorially complex but with a relatively larger 'intuitive loading' on the illustrative dimension. The complex illustrations may make the dimensions less distinct and serve to reduce their discriminant validity" (Dickinson & Tice, 1977, pp. 218-219).

Research concerning factor analysis of BARS data

Few empirical tests of these questions exist in the research literature. Dickinson and Tice (1977) screened out factorially complex behavioral items in a BARS measuring firefighter performance by using a multiple group factor
analysis of the scale values of all items. A checklist format was employed in which the raters simply checked behavioral items typical of ratee behavior, and the scale value of these items was summed to obtain a score for each firefighter on each of three dimensions. The factor analysis indicated that of the original set of 40 behavioral items, 15 were factorially complex. A behavioral item was eliminated if its loading on the hypothesized factor (the intuited performance dimension) was less than .40 or if less than 85 percent of its communality fell on the hypothesized factor. The 25 remaining items were then compared to the original 40 in order to check halo levels. Subjecting the data to an ANOVA, the 25 screened items yielded ratings with greater dimension independence than the 40 unscreened items. The authors suggested that raising the criteria for items in the retranslation step (for instance, from 60 percent to 85 percent agreement) may help to screen out factorially-complex items and thereby reduce halo.

In a study in which behavioral items critical to the job success of logging supervisors were rated on a 5-point Likert scale by two independent sets of observers (Latham & Wexley, 1977), a factor analysis reduced 78 behavioral items to 38 and 33 for the two sets of raters. These items constituted 10 and 11 factors, respectively, which in turn were used for performance evaluation purposes. The reliability and relevance (as measured by levels of concurrent validity) of these factor-analytically derived scales were
then compared to the intuitively-derived scales from which
the behavioral items originated (Latham, Wexley, & Rand,
1975, as cited in Latham & Wexley, 1977). The results of
this study indicated that the scales developed by factor
analyzing ratings had moderately higher reliability and ac-
counted for as much, if not more variance in the cost-related
factors (productivity, absenteeism, and attendance) as the
scales developed by the traditional intuitive method. The
authors suggested that those findings may have been due to
the more efficient use of information in the item pool as
a result of the screening of ambiguous or redundant items.

In an experiment examining BARS' psychometric prop-
erties in a field setting, Kavanagh and Duffy (1978) analyzed
data obtained from ratings of an educational television se-
ries designed to teach children reading skills. These rat-
ings were subjected to a pair of factor analyses in order
to check if "descriptors" (anchors in this case were de-
scriptions of film segments) clustered together on the same
dimensions as they did in scale development. Only a modest
amount of agreement was found, with the items from only one
intuited dimension loading together on the factor analysis.
This five-factor solution accounted for 56.9 percent of the
total variance. Another factor analysis was conducted, this
time with eight factors extracted, as suggested by the
"eigenvalue drop-off" criterion (see Harmon, 1967). This
eight-factor solution accounted for 67.5 percent of the var-
iance and yielded a more interpretable factor structure than
the five-factor solution. These findings are contrary to the hope held by Smith and Kendall that field use would reduce rather than expand the number of dimensions. However, one purpose of the retranslation technique -- to better define the scale dimension -- seems to have been met with the increased number of dimensions (Kavanagh & Duffy, 1978).

Analyzing the data from four independent samples of BARS applications across three different occupations (professors, police officers, and nurses) Kafry, Jacobs and Zedeck (1979) examined a series of BARS in an attempt to point out an appropriate number of dimensions to be included in such studies. The appropriate number of dimensions was defined here as "non-redundant dimension information" (p. 188). The focus was to reduce the total number of rated dimensions while retaining as much of the original information as possible. The number of original dimensions in each study ranged from nine to twenty-four. The results of the factor analysis showed that less than one-half the number of original dimensions could explain approximately 90 percent or more of the total dimension variance in each of the four studies. Across these studies, the number of dimensions necessary to explain 90 percent of the total variance was reduced between 56 percent (from nine dimensions to four) and 63 percent (twenty four dimensions to nine). Kafry et al. suggested that scale developers should "limit the number of rated dimensions to be consistent with both the minimum number requisite for adequate job description and the maximum
number of discriminable stimuli for a group of raters" p. 190). To determine whether high scale intercorrelations are a function of an excessive number of dimensions overloading the cognitive capacity of the rater, the authors suggested that future researchers should examine whether dimension independence will be a function of individual differences in the level of the cognitive complexity of the raters.

**Implications of factorial complexity within BARS**

Schneier (1977) examined the question of cognitive complexity by having 60 manufacturing workers rate the performance of their peers using either a traditionally developed BARS or a much simpler scale requiring far fewer discriminations (this scale contained ten dimensions to be rated with the adjectives "above average," "average," and "below average"). All raters were identified as cognitively complex or cognitively simple by virtue of their performance on the Bieri grid form (see Bieri, Atkins, Briar, Leaman, Miller, & Tripodi, 1966, pp. 190-191). Schneier hypothesized that cognitively complex raters would exhibit less haloed ratings when using the more complex rating form (BARS) than when using the simpler scales. Results confirmed that the cognitively complex raters did have less halo in their ratings when using the BARS format, but contrary to expectations, they also showed lower halo levels than the cognitively simple group when using the simplified rating scales. Cognitively simple raters showed lower halo levels in their ratings when they used the simplified rating scale. When the
complexity of the rater and the format were matched, raters felt more confident in their ratings and preferred the format they were using. Thus, cognitive complexity seems to interact with rating format to produce ratings exhibiting halo error. Schneier suggests that the cognitive complexity of the rater ought to be incorporated into future research designs in order to specify those raters for which BARS may be most appropriate.

The research cited above seems to suggest that the halo levels present in BARS ratings may be moderated by the cognitive characteristics of raters. While the majority of BARS research has focused on various aspects of the developmental process in hopes of finding methods to reduce halo, differences in cognitive levels of raters (and hence their ability to effectively use BARS) may be partly responsible for the ubiquitous nature of halo error. The large number of discriminations required of the rater may overmatch their cognitive abilities, and therefore he/she may simply rely more on a global impression of the ratee which causes the ratings to be haloed (Schneier, 1977).

Although subsequent research has called into question the validity of measures of cognitive complexity (Lahey & Saal, 1981; Sauser & Pond, 1981; Bernardin, Cardy & Carlyle, 1982), this concept remains consistent with two of the sources of halo (engulfing and cognitive distortions) listed by Cooper (1981a). In both sources, information available to the rater's implicit belief about the degree to which
dimensions co-vary. Engulfing occurs when raters can no longer make the necessary distinctions between dimensions and rely on their own beliefs regarding dimension covariance in making their ratings. The cognitive distortion process is one in which information stored in the memory of the rater is lost, and again the rater relies on his own notions of dimension covariance to make his ratings. The rater may be forced into such a situation where the number of dimensions (and thus the number of discriminations required) over-match his/her cognitive abilities.

There also exists the possibility that the ubiquitous nature of halo may be the result of "factorially complex" behavioral incidents anchoring BARS. The intuitive factor analysis performed by scale developers may result in anchors which tap more than one performance dimension. This may confuse the rater and reduce the distinctiveness of the separate dimensions (Dickinson & Tice, 1977), thereby increasing the risk of halo. While the intuition of dimensions by raters increases their participation in scale development, it may not be sound psychometric practice since it allows some of the factorially-complex behavioral incidents to blur the dimensions and increase halo.

A possible method of obtaining "purer" and more distinct performance dimensions might be to subject ratings on the individual behavioral incidents (across all performance dimensions) to a factor analysis where these incidents may be sorted into factors which are more distinct than the dimensions
intuited by raters during BARS development. The results of this factor analysis will be used to redistribute (or possibly eliminate) some of the incidents among the resulting scales, thereby creating a modified version of the original BARS. Theoretically, this process should serve to make dimensions more independent and less susceptible of halo. A factor analysis may also reduce the total number of rated dimensions, thus reducing the cognitive burden on the rater and allowing him to make the required discriminations with greater ease. Of course, the possibility also exists that the factor analysis may indicate that a larger number of factors is appropriate. Regardless of the resulting number of dimensions, the goal of the procedure is to group observable job behaviors into rating dimensions in such a way that the rater is more likely to perceive the distinctiveness of the dimensions and thus reduce halo.

In the research conducted here, ratings were collected on the individual behavioral anchors composing a BARS designed to measure the teaching effectiveness of a university psychology instructor (Harari & Zedeck, 1973). These ratings were factor analyzed with the intention of developing a set of scales with more independent dimensions of performance. It is hoped that these modified scales will be less cumbersome for the rater to use and will provide more distinct dimensions resulting in halo levels lower than the original form of BARS.
METHOD

Overview

The purpose of this study was to compare the amount of halo error present in an original set of BARS (Harari & Zedeck, 1973, developed for the purpose of rating university psychology instructors) with halo found in a modified version of these scales. The dimensions of job performance composing the modified BARS were those identified through a factor analysis of ratings on the individual behavioral anchors taken from the Harari-Zedeck BARS. The halo levels found in the modified BARS and the Harari-Zedeck BARS were compared using interdimension correlations as indicators of halo.

Subjects

Two hundred thirty-six students enrolled in undergraduate psychology courses at Western Kentucky University participated as raters in this study. Seven members of the psychology department faculty (five males and two females) were ratees.

Procedure

The research design of this study required that three distinct forms of BARS be presented to raters over two phases of data collection. In Phase 1 the individual behavioral
anchors taken from the Harari-Zedeck BARS were randomly ordered and presented to the raters with no indication of the dimension to which they belonged. In Phase 2 students rated their instructors' performance using either the Harari-Zedeck BARS or a modified version of these scales. It was in Phase 2 that halo levels were compared between the two BARS. In all three forms of BARS the wording of the behavioral anchors was identical.

Phased 1. In this initial phase of data collection raters were asked to respond to each of the 79 Harari-Zedeck behavioral anchors (presented in random order with no accompanying dimension title or definition) using a Likert-type scale. For example, in the dimension labelled "Interpersonal Relations with Students" one of the behavioral anchors reads "This professor could be expected to try to humiliate students who disagree with him." Raters judge the likelihood of the professor's displaying that particular behavior using a 1-7 Likert scale with the term "very unlikely" and "very likely" anchoring the low end and the high end of the continuum, respectively. The raters were not informed to which dimension a behavioral anchor belonged since anchors were randomly ordered and presented without dimension titles or definitions. The rater provided similar judgements on the remaining anchors until all 78 were rated. Details of the composition of the Harari-Zedeck BARS appear below in Phase 2.

After all rating forms had been collected from students, the ratings on the behavioral anchors were intercorrelated
and the resulting 78 x 78 correlation matrix subjected to a common factors analysis. In determining the proper number of factors to be extracted in the factor analysis, both Catell's scree test (1966) and Kaiser's criterion were employed. Catell's scree test uses a graph in which eigenvalues are plotted against their corresponding factors (in the order in which the factors are extracted) and the shape of the resulting curve determines the cutoff point. Specifically, the point at which the curve's slope changes most rapidly is taken as the maximum number of factors to be extracted (Child, 1970). A currently more popular method of making this determination is the use of Kaiser's criterion. In this method only eigenvalues greater than 1.00 are considered as common factors. However, Catell has suggested that this criterion is most appropriate when the number of variables is between 20 and 50. When more than 50 variables are involved (there were 78 in this study) there is a tendency for Kaiser's criterion to extract too many factors (Child, 1970). Since one of the stated goals of this study was to reduce the number of rated dimensions, it was decided that Catell's scree test would be used in conjunction with Kaiser's criterion to determine the proper number of factors to be extracted.

The retained factors were rotated to Thurstone's simple structure using the varimax procedure. This rotational scheme was chosen over other orthogonal rotations because varimax aids in this interpretation of each factor (as opposed to
procedures which seek to interpret the underlying variables). According to Weiss (1976), varimax is the proper orthogonal rotation when the objective is to interpret the underlying factors or to understand the factor composition of a set of variables.

The purpose of the factor analysis of ratings collected in Phase 1 was to guide the development of a new factor analytically-derived BARS, which was to be compared with the Harari-Zedeck BARS in Phase 2. Three criteria were used to decide if an incident should be included in the modified BARS. First, the rotated factor matrix was examined to check each incident's factor loadings across all factors. Only incidents loading greater than .30 for a given factor were considered for inclusion. Secondly, if the difference between an incident's highest and next highest factor loading was less than .10, that incident was dropped from further consideration. This methodology follows the precedent set by Dickinson and Tice (1977) who screened out such incidents on the basis of being "factorially complex" (p. 218). Finally, all incidents within a given factor passing the first two criteria were screened in order to determine whether they were conceptually similar. For example, if an item seemed to be measuring the instructor's ability to motivate students while the other incidents within that factor referred to the instructor's grading policies, that incident would be dropped from the factor. Any factors which contained a wide variety of behavioral anchors which seemed to measure
more than one dimension of performance were eliminated completely. Based upon these considerations, a modified version of BARS was formed containing anchors from the Harari-Zedeck BARS redistributed among new factors.

Phase 2. In this phase of the study a new sample of raters was asked to rate their instructors' performance using either the original Harari-Zedeck BARS or the modified BARS (developed from ratings collected in Phase 1). Harari and Zedeck followed the BARS development prescriptions set forth by Smith and Kendall (1963) in developing their BARS used to rate the teaching performance of university psychology instructors. Harari and Zedeck identified nine dimensions relevant to teaching performance: 1) Depth of Knowledge, 2) Delivery, 3) Organization, 4) Interpersonal Relations with Students, 5) Relevance, 6) Testing, 7) Grading, 8) Assignments and Workload, and 9) Ability to Motivate. For a full description of scale development and statistics concerning each of these dimensions, refer to Harari and Zedeck (1973). These scales are presented in their complete form in Appendix A. The modified BARS was based upon the results of the factor analysis of ratings collected in Phase 1. The dimensions of job performance composing the modified BARS were those identified through the factor analysis. Both the Harari-Zedeck BARS and the modified BARS appeared in a typical graphic BARS format.

Raters in Phase 2 were 155 undergraduate psychology students. Five faculty members of the psychology department
served as ratees. Raters were randomly assigned to one of
two rating conditions, using either the Harari-Zedeck BARS
or the modified BARS to rate their instructor's performance.
The order in which the rated dimensions appear were random-
ized in each BARS condition. A comparison was made in order
to test for differences in halo levels between the two sets
of scales, as measured by the interdimension correlations
of each.

Raters in both conditions were asked to read the dimen-
sion titles and definitions of the scales before making any
judgement as to where along the continuum an instructor's
performance fell. Raters were then asked to cite two exam-
ple behaviors typical of their instructor's performance
relevant to the dimension being rated, and to place a mark
along the continuum at the point which the rater feels is
representative of the levels of effectiveness for each of
the two documented behaviors. The numerical average of these
two points is taken as the instructor's "score" for the dimen-
sion.

Smith and Kendall (1963) believed that having raters
document observed behaviors and place them along the rating
continuum was important for two reasons. First, they wanted
to impress upon the rater that the behavioral anchors were
only reference points placed along the continuum to give
him a frame of reference for the effectiveness of certain
behaviors. The rater was not asked to search the list of
anchors until he found one that his instructor had actually
exhibited. Secondly, the documentation of behaviors observed by the rater permitted the checking and possible revision of the scales.

Following the collection of all rating forms, inter-dimension correlations were calculated across ratees and then converted to standard scores using Fisher's Z transformation. A t-test for unequal N's was used to determine if a significant difference existed between the two BARS' inter-dimension correlations. The form of BARS having lower inter-dimension correlations contained the lesser amount of illusory halo.
RESULTS

Phase 1

Ratings from all 78 behavioral anchors were intercorrelated and the resulting correlation matrix subjected to a common factors analysis. Neither Cattell's scree test nor Kaiser's criterion (for determining the proper number of factors to extract in a factor analysis) provided an interpretable factor structure. Cattell's scree test suggested that five factors should be extracted, but upon examining this five factor solution, each factor contained a wide variety of behavioral anchors, and none seemed to tap a definable dimension of teaching performance. Kaiser's criterion (in which eigenvalues greater than 1.00 are considered as common factors) suggested that 25 factors be extracted. Kaiser's criterion was obviously not a feasible alternative since such a large number of factors would be counterproductive in the attempt to produce a more distinct and less cumbersome set of scales. This supports Cattell's warning that in factor analyses involving more than 50 variables, there is a tendency for Kaiser's criterion to extract too many factors.

It was then decided to specify a nine-factor solution (corresponding to the number of performance dimensions in the original BARS) and use only those anchors and factors
meeting the criteria mentioned previously (see pp. 44-45 of Method). The initial screening criterion was the elimination of any item loading less than .30 on all factors. It was felt that this criterion was needed to eliminate any poorly written or conceptually incompatible items appearing in the scales. Only five of 79 items were eliminated on this basis. Thus, it can reasonably be assumed that the conventional BARS development process is useful for screening out these incompatible behavioral items.

The second screening criterion involved comparing the highest and next-highest factor loading for each behavioral item. If this difference was less than .10, the item was eliminated. This criterion was necessary to screen out any items which may have been "factorially complex," or conceptually related to more than one dimension of teaching performance. This problem is what Dickinson and Tice (1977) believed to be responsible for haloed performance dimension ratings and reduced discriminant validity.

In the present study, over one-third (25 of 74) of the behavioral items were eliminated on the basis of this second criterion. We can thus assume a sizable portion of the behavioral items appearing in a conventional BARS to be representative of at least two supposedly distinct dimensions. It appears that the conventional BARS development process is not nearly as efficient in eliminating "factorially complex" items as it is in screening conceptually incompatible items.
Of the remaining 49 behavioral items, approximately one-half (24) were eliminated on the basis of the final screening criterion. The final criterion involved a subjective review of the items within a given factor and the elimination of those which did not reflect the general content of the other items. This resulted in the elimination of two entire factors and the nine behavioral items which composed them. This elimination made in the belief that any factor composed of a heterogenous blend of behavioral items was clearly factorially complex, and thus likely to contribute to halo error.

Employing these criteria, a seven-factor solution was found to be the most interpretable (see Table 1). Based upon a post hoc examination of the content of the anchors composing each of the factors, these newly derived dimensions were labeled 1) Ability to Motivate Student, 2) Handling Questions and Class Discussion, 3) Delivery, 4) Knowledge and Application of Subject Material, 5) Testing, 6) Organization, and 7) Grading. The number of anchors in these dimensions ranged from three to five. The definition of each dimension was designed to reflect the collective content of the behavioral anchors it contained. These modified scales are presented in their complete form in Appendix B.

Next, both the Harari-Zedeck and the modified BARS were presented in their complete forms to a new set of raters in Phase 2 of the study. The results of this phase appear below.
Phase 2

In this final phase of data collection, 155 undergraduate psychology students were asked to rate their instructor's teaching performance using either the Harari-Zedeck BARS or the modified BARS. Five members of the psychology department faculty served as ratees.

As noted previously, raters were asked to provide two relevant examples of their instructor's behavior for each rated dimension. Due to time constraints, raters were not always able to provide two examples. Raters using the Harari-Zedeck BARS listed an average of 1.46 behaviors per dimension, and those using the modified BARS listed an average of 1.44 behaviors. Where two behaviors for a single dimension were reported and scaled, the numerical average was taken as the instructor's "score" for that dimension.

Of the 155 raters participating in this phase, 85 (54.8 percent) used the modified BARS to rate their instructor. A total of 70 raters (45.2 percent) used the Harari-Zedeck BARS. All dimension scores were correlated across raters within forms of BARS. The resulting correlation matrices can be found in Tables 2 and 3.

A t-test for unequal N's was used to compare the two BARS' interdimension correlations. A total of 36 non-redundant intercorrelations taken from the matrix of the Harari-Zedeck BARS were compared with 21 non-redundant intercorrelations taken from modified BARS matrix. The difference between the two sets of correlations was not significant, \( t (55) = 1.017, p \approx .05 \).
Contrary to the expected results, the mean of the intercorrelations was slightly higher for the modified BARS (x = .327) than for the Harari-Zedek BARS (x = .282), although this difference was not statistically significant. For the Harari-Zedek BARS, the intercorrelations ranged from a high of .525 (Organization-Depth of Knowledge) to a low of .003 (Ability to Motivate-Grading). The intercorrelations among the dimensions of the modified BARS ranged from .516 (Delivery-Testing) to -.010 (Testing-Grading).

As mentioned earlier in this text, the possibility exists that reduced halo levels may in some cases be due to a reduction in scale reliability. To investigate this possibility, a post hoc analysis was conducted in which the interrater reliabilities of each form of the BARS were calculated and compared. For each of the seven classes in which both forms of BARS were used to rate teaching performance, a separate interrater reliability coefficient was calculated. These seven pairs of coefficients were then compared using a t-test for matched pairs to check whether the Harari-Zedek and modified BARS were significantly different in terms of scale reliability. Results showed that the Harari-Zedek scales had significantly lower reliabilities, t (6) = 3.1872, p < .05. The reliabilities are reported by individual class section in Table 4.
DISCUSSION

The objective of the present study was the development of a modified set of behaviorally anchored rating scales which were less cumbersome to use and with halo levels lower than those found in conventional BARS. It was hypothesized that a factor analysis of ratings obtained on the individual behavioral incidents would lead to the development of scales having lower levels of halo, as measured by the magnitude of the interdimension correlations. Despite the fact that these factor analytically-modified scales contained two fewer dimensions and 54 fewer behavioral incidents than the conventional Harari-Zedeck BARS, the hypothesized results did not occur.

Before addressing the issue of the modified BARS' failure to provide less haloed ratings, a discussion of the results of Phase 1, the portion of the study in which the modified BARS was constructed, is presented.

Phase 1

As noted previously, both Cattell's scree test and Kaiser's criterion proved to be ineffective in suggesting the proper number of factors to be extracted. The five-factor solution suggested by Cattell's scree test was not easily interpretable, and the 25-factor solution derived
by Kaiser's criterion was much too unwieldy to be of practical use. Consequently, a nine-factor solution was specified since the original Harari-Zedeck BARS contained nine dimensions. The factors obtained from this solution were far more cohesive and interpretable than those provided by previous solutions. Each behavioral item was tentatively assigned to one of nine factors based upon the item's highest factor loading. The criteria referred to on pages 44-45 of the Method were then employed to "pare down" both the number of individual anchors and factors.

The last of these screening criteria (the inspection of the items within a given factor for conceptual similarity) is obviously a subjective procedure. Items having factor loadings greater than .6000 were sometimes eliminated on the basis of apparent conceptual incompatibility with the content of the remaining items. Since this study was intended to empirically (rather than intuitively) construct independent scales, in retrospect it might have been wiser to choose behavioral items for a given dimension based strictly upon the item's highest factor loading rather than attempt to select only those which seemed to be conceptually similar. Perhaps the eliminated items related to their dimensions in a manner not apparent to the researcher, and thus should not have been removed. However, it was felt that a "common sense" approach in the form of the subjective removal of some items would be useful in improving the psychometric quality of the scales. Since this subjective approach to
the final screening criterion was not in strict keeping with
the empirical nature of the modified BARS development process,
future researchers may wish to drop this final criterion
when attempting to develop factor analytically-derived scales.

In combination, the criteria used to screen the beha-
vioral incidents was expected to produce a modified set of
BARS which would be characterized by lesser amounts of illu-
sory halo than the original Harari-Zedeck BARS. Since this
hypothesis was not confirmed, a review of Phase 2 of the
study might help to explain these findings.

**Phase 2**

Although in general the dimension intercorrelations
among the Harari-Zedeck and the modified BARS were not signi-
icantly different, a closer look at the correlation matrices
reveals some interesting findings. Among the dimensions
of the Harari-Zedeck BARS, "Depth of Knowledge" had the
highest mean interdimension correlation (\( r = .354 \)), followed
by "Testing" (\( r = .343 \)) and "Ability to Motivate" (\( r = .309 \)).
Thus, the instructor's knowledge of the subject material
was judged by raters to be the area of performance most close-
ly related to overall teaching effectiveness. This seems
to make intuitive sense in a situation where a student has
limited exposure to a given instructor (the majority of raters
in the present study were introductory psychology students).
The instructor's knowledge of his/her topic and the ability
to generate student interest (Ability to Motivate) might
appear to be the most salient aspects of performance to a
student having limited exposure to a given instructor. The relatively strong mean interdimension correlation of Testing was somewhat of a surprise, especially in view of its extremely weak correlations within the modified BARS (an issue which will be addressed below). It is not surprising, however, that the Testing dimension correlated most highly with Grading (.523) and Assignments and Workload (.460). All three of these dimensions are relatively concrete, observable areas of performance with which all students have familiarity.

In general, the intercorrelations among the dimensions of the modified BARS were slightly higher than those of the Harari-Zedeck BARS, an unexpected result in this study. Within the modified BARS, Ability to Motivate Student, Organization, and Knowledge and Application of Subject Material were the dimensions with the highest mean interdimension correlations. These dimensions are similar in terms of their content to those dimensions in the Harari-Zedeck BARS also found to have high mean interdimension correlations. The dimension Knowledge and Application of Subject Material is highly similar to the dimension within the Harari-Zedeck BARS labeled "Depth of Knowledge." (It should be noted here that since the modified BARS was derived via factor analysis from the Harari-Zedeck Bars, several of the dimension titles and definitions are identical and others are similar. The reader should refer to Appendices A and B to avoid confusion). In the modified BARS, it again becomes apparent that the instructor's ability to generate student enthusiasm and his/her
knowledge of the field are judged to be the most strongly related to overall teaching effectiveness. The instructor's ability to present material and conduct classes in an organized manner is also seen as relevant to successful teaching. Since the range of the mean interdimension correlations in both sets of scales was fairly restricted, no firm conclusions can be safely drawn regarding those with the highest mean intercorrelations.

Performance dimensions having low correlations with other dimensions also provided interesting data. Within the Harari-Zedek BARS, the dimensions labeled "Interpersonal Relations" and "Grading" had the lowest mean intercorrelations, .226 and .231 respectively. While this seems to suggest that personal relationships between the instructor and students along with classroom grading policies are not of utmost relevance to overall teaching ability, the differences among dimension intercorrelations are not great enough to warrant any conclusive statements. It is interesting to note, however, that three near-zero intercorrelations occurred within the Grading dimension of the Harari-Zedek BARS. "Grading" and "Ability to Motivate" correlated .003, while Grading-Delivery correlated .070 and Grading-Organization correlated .071. Raters apparently saw virtually no connection between an instructor's grading policies and his/her ability to motivate students.

Within the dimensions of the modified BARS, the mean interdimension correlations were relatively homogenous, with
the striking exception of the Grading dimension \((x = .070)\). This dimension correlated no higher than \(.152\) (Grading-Organization) with any of the remaining six dimensions, and contained three near-zero correlations (Grading-Handling Questions and Class Discussion = .016; Grading-Delivery = .009; Grading-Testing = -.010). The correlation between Grading-Ability to Motivate (.003 in the Harari-Zedeck BARS), was again found to be quite low, correlating \(.119\) in the modified BARS. Raters using the modified BARS apparently saw little relationship between an instructor's grading policies and all other aspects of teaching performance.

In review, it appears that in both scales dimensions referring to the instructor's mastery of the subject matter and his/her ability to motivate students are seen as the most closely related to overall effectiveness (as measured by the relative magnitude of interdimension correlations). Thus in terms of halo research, these are the areas in which scale developers need to exercise the most caution in order to limit illusory halo. Future research into BARS development may reveal how much of this relationship is illusory halo and how much is true halo.

Shortcomings of the modified BARS

The reader will recall that one of the proposed advantages of the modified BARS was the development of a set of scales with fewer dimensions and behavioral anchors, thus reducing the required number of discriminations the scale user must make (Kafry, Jacobs, & Zedeck, 1979; Schneier,
1977). The modified BARS in the present study reduced the number of rated dimensions from nine (in the Harari-Zedock BARS) to seven and reduced the number of behavioral anchors by nearly 70 percent (from 79 to 25). If Kafry et al. and Schneier were correct in assuming that a conventional BARS overloads a typical rater's capacity for making discriminations among stimuli, this factor analytically-reduced BARS should enable the rater to make quicker and easier decisions, theoretically characterized by reduced levels of illusory halo. Since halo levels were not reduced (in fact there was a slight increase in halo), skepticism regarding cognitive complexity and the usefulness of reducing the number of discriminations is well founded. Nearly the same number of newly scaled incidents were elicited from raters in both conditions despite the greater number of dimensions and anchors confronted by raters using the Harari-Zedeck BARS. Thus, the modified BARS appeared to be no quicker or easier for the rater to use.

One problem typically encountered in BARS development is a lack of reliable behavioral anchors near the scale midpoint. This problem is a result of the criteria prescribed by Smith and Kendall (1963) in the fourth step (scaling incidents) of the conventional BARS developmental process. In this step items are screened based upon the standard deviation of their scale value as judged by a group of potential raters. In most studies (see DeCotiis, 1978, p. 683) a standard deviation of less than 1.5 is required for the item
to pass this screening criterion. Since it is fairly common
for mid-range scale items to have relatively higher standard
deviations, a problem often develops in finding mid-scale
items which pass this criterion, thus creating a gap in this
case of the scale continuum. This problem was expected to
be exaggerated during the development of the modified BARS
as a result of the additional criteria imposed. It was
feared that illusory halo would be increased, since raters
might be tempted to "fill in" the gaps with their newly-
written behaviors. In order to investigate this potential
problem, the portion of each scale continuum where these
gaps occurred (i.e., areas devoid of behavioral anchors)
was examined in order to check if they contained a disprop-
portionate number of newly written behaviors. A ratio was
computed in which the size of the scale gap for each dimen-
sion was divided by the overall length of the scale. This
ratio was compared to the proportion of newly scaled beha-
viors placed within these gaps by all raters. This process
was repeated for all seven dimensions within the modified
BARS. This review of the data revealed that raters scaled
behaviors within these gaps with no greater frequency than
could normally be expected. Therefore, these scale gaps
cannot be blamed for the failure to reduce halo levels.

A possible explanation for the failure of the modified
BARS to reduce halo levels may have been the starting point
of its development. Essentially, its development began with
the finished product of the Harari-Zedeck BARS, at least
in terms of the final pool of behavioral items. The modified BARS began with a pool of items which had already been reduced from 310 to 79 by the conventional BARS process. A better approach might have been to begin the process with the original pool of 310 items elicited in the initial phase of Harari-Zedeck's study, and have student/raters rate their instructors using these items (as in Phase 1 of the present study). The modified BARS would then be formed from the results of this factor analysis. Using a factor analysis rather than the retranslation method to eliminate items might result in a more efficient screening of factorially-complex items. Disadvantages would be the increased time required in the initial rating phase (Phase 1 of the present study would have taken nearly four times longer) and decreased rater participation in the development process. Raters are still highly involved in the initial process however, and as pointed out by Jacobs, Kafry and Zedeck (1980), participation may not be essential to the reduction of halo.

Another possible explanation for the modified BARS' failure to demonstrate lower halo levels than the Harari-Zedeck BARS may rest with the interrater reliabilities of each. As reported earlier in this text (see p. 24), in some cases reduction of halo can be traced to decreased scale reliability. Since interrater reliability was significantly lower for the Harari-Zedeck BARS than the modified BARS in this study, the possibility exists that lower scale reliability was partly responsible for reducing halo levels in
the Harari-Zedeck BARS. This statistical artifact may thus be responsible for the lack of significant differences in halo levels among the two sets of scales. Future researchers may wish to examine the possibility that a factor-analytic approach to BARS development may result in improved scale reliability over the conventional BARS development process.

Recommendations for the future use of BARS

For the past 15 years, a tremendous amount of research has been conducted on the psychometric properties of BARS and their potential for reducing levels of illusory halo. The present study is yet another addition, and as the majority of BARS/halo research, the attempt to reduce levels of halo was not successful. These findings do not mean, however, that BARS have no practical use. BARS have all the appearances of being an elegant, well-thought-out tool that simply does not offer any real improvements to rating systems already in use. As Blood (1974) pointed out over a decade ago, this does not necessarily spell the end of BARS use. He suggested that the behavioral items collected in the development process would provide an effective basis for a training program. Skills relevant to job success are specified in terms of the actual job behaviors rather than a vaguely related area of performance. The research conducted by Zedeck and Baker (1972) showed that the scaling phase of BARS development (where the scale values of behavioral items are established) can reveal job areas where different levels of the
organizational hierarchy (i.e., managers versus sales clerks) are in disagreement over the value of particular job behaviors. Rated job behaviors with large variances identify areas where organizational policy is unclear or in conflict (Blood, 1974).

Perhaps because of BARS' intuitive appeal as a performance rating device, other practical uses have been ignored. Perhaps the time has arrived when applied research should focus on these additional uses.
Table 1
Behavioral Incidents Composing Modified BARS

<table>
<thead>
<tr>
<th>Item #</th>
<th>Dimension Title</th>
<th>Factor Loading</th>
<th>Harari-Zedeck Dimension From Which Item Orig.</th>
<th>Scale Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>42</td>
<td>Ability to Motivate</td>
<td>.61519</td>
<td>Ability to Motivate</td>
<td>5.3</td>
</tr>
<tr>
<td>26</td>
<td>Ability to Motivate</td>
<td>.55240</td>
<td>Ability to Motivate</td>
<td>6.5</td>
</tr>
<tr>
<td>5</td>
<td>Ability to Motivate</td>
<td>.53125</td>
<td>Ability to Motivate</td>
<td>6.0</td>
</tr>
<tr>
<td>9</td>
<td>Ability to Motivate</td>
<td>.50365</td>
<td>Delivery</td>
<td>1.5</td>
</tr>
<tr>
<td>13</td>
<td>Handling Questions</td>
<td>.64620</td>
<td>Interpersonal Relations</td>
<td>6.2</td>
</tr>
<tr>
<td>38</td>
<td>Handling Questions</td>
<td>.58067</td>
<td>Interpersonal Relations</td>
<td>1.3</td>
</tr>
<tr>
<td>76</td>
<td>Handling Questions</td>
<td>.53646</td>
<td>Depth of Knowledge</td>
<td>2.0</td>
</tr>
<tr>
<td>20</td>
<td>Handling Questions</td>
<td>.52543</td>
<td>Relevance</td>
<td>2.1</td>
</tr>
<tr>
<td>29</td>
<td>Handling Questions</td>
<td>.40300</td>
<td>Depth of Knowledge</td>
<td>3.1</td>
</tr>
<tr>
<td>25</td>
<td>Delivery</td>
<td>.51911</td>
<td>Delivery</td>
<td>1.8</td>
</tr>
<tr>
<td>24</td>
<td>Delivery</td>
<td>.48839</td>
<td>Delivery</td>
<td>3.6</td>
</tr>
<tr>
<td>19</td>
<td>Delivery</td>
<td>.41783</td>
<td>Delivery</td>
<td>3.8</td>
</tr>
<tr>
<td>62</td>
<td>Knowledge &amp; Application</td>
<td>.55224</td>
<td>Relevance</td>
<td>6.1</td>
</tr>
<tr>
<td>15</td>
<td>Knowledge &amp; Application</td>
<td>.54723</td>
<td>Relevance</td>
<td>2.6</td>
</tr>
<tr>
<td>34</td>
<td>Knowledge &amp; Application</td>
<td>.45314</td>
<td>Depth of Knowledge</td>
<td>2.8</td>
</tr>
<tr>
<td>39</td>
<td>Knowledge &amp; Application</td>
<td>.42563</td>
<td>Depth of Knowledge</td>
<td>3.8</td>
</tr>
<tr>
<td>1</td>
<td>Testing</td>
<td>.57363</td>
<td>Testing</td>
<td>3.1</td>
</tr>
<tr>
<td>67</td>
<td>Testing</td>
<td>.47675</td>
<td>Testing</td>
<td>3.7</td>
</tr>
<tr>
<td>3</td>
<td>Testing</td>
<td>.46814</td>
<td>Testing</td>
<td>2.4</td>
</tr>
<tr>
<td>41</td>
<td>Organization</td>
<td>.47300</td>
<td>Organization</td>
<td>5.5</td>
</tr>
<tr>
<td>44</td>
<td>Organization</td>
<td>.46517</td>
<td>Organization</td>
<td>5.8</td>
</tr>
<tr>
<td>55</td>
<td>Organization</td>
<td>.39245</td>
<td>Assignments &amp; Workload</td>
<td>5.1</td>
</tr>
<tr>
<td>72</td>
<td>Grading</td>
<td>.66406</td>
<td>Grading</td>
<td>3.2</td>
</tr>
<tr>
<td>77</td>
<td>Grading</td>
<td>.60615</td>
<td>Grading</td>
<td>2.2</td>
</tr>
<tr>
<td>69</td>
<td>Grading</td>
<td>.44555</td>
<td>Grading</td>
<td>5.4</td>
</tr>
<tr>
<td>MOTIV</td>
<td>DELIV</td>
<td>KNOWL</td>
<td>RELEV</td>
<td>RELAT</td>
</tr>
<tr>
<td>-------</td>
<td>-------</td>
<td>-------</td>
<td>-------</td>
<td>-------</td>
</tr>
<tr>
<td>---</td>
<td>0.349</td>
<td>0.327</td>
<td>0.478</td>
<td>0.160</td>
</tr>
<tr>
<td>DELIV</td>
<td>---</td>
<td>0.356</td>
<td>0.124</td>
<td>0.237</td>
</tr>
<tr>
<td>KNOWL</td>
<td>---</td>
<td>---</td>
<td>0.349</td>
<td>0.246</td>
</tr>
<tr>
<td>RELEV</td>
<td>---</td>
<td>---</td>
<td>---</td>
<td>0.230</td>
</tr>
<tr>
<td>RELAT</td>
<td>---</td>
<td>---</td>
<td>---</td>
<td>---</td>
</tr>
<tr>
<td>WORKL</td>
<td>---</td>
<td>---</td>
<td>---</td>
<td>---</td>
</tr>
<tr>
<td>TESTI</td>
<td>---</td>
<td>---</td>
<td>---</td>
<td>---</td>
</tr>
<tr>
<td>ORGAN</td>
<td>---</td>
<td>---</td>
<td>---</td>
<td>---</td>
</tr>
<tr>
<td>GRADI</td>
<td>---</td>
<td>---</td>
<td>---</td>
<td>---</td>
</tr>
</tbody>
</table>

**MOTIV = ABILITY TO MOTIVATE**  
**DELIV = DELIVERY**  
**KNOWL = DEPTH OF KNOWLEDGE**  
**RELEV = RELEVANCE**  
**RELAT = INTERPERSONAL RELATIONS WITH STUDENTS**  
**WORKL = WORKLOAD AND ASSIGNMENTS**  
**TESTI = TESTING**  
**ORGAN = ORGANIZATION**  
**GRADI = GRADING**
### Table 3

**Dimension Intercorrelations for the Modified BARS**

<table>
<thead>
<tr>
<th></th>
<th>MOTIV</th>
<th>HANDL</th>
<th>DELIV</th>
<th>APPLI</th>
<th>TESTI</th>
<th>ORGAN</th>
<th>GRADI</th>
</tr>
</thead>
<tbody>
<tr>
<td>MOTIV</td>
<td>---</td>
<td>0.416</td>
<td>0.504</td>
<td>0.474</td>
<td>0.397</td>
<td>0.412</td>
<td>0.119</td>
</tr>
<tr>
<td>HANDL</td>
<td>---</td>
<td>---</td>
<td>0.234</td>
<td>0.437</td>
<td>0.273</td>
<td>0.285</td>
<td>0.016</td>
</tr>
<tr>
<td>DELIV</td>
<td>---</td>
<td>---</td>
<td>---</td>
<td>0.416</td>
<td>0.516</td>
<td>0.441</td>
<td>0.009</td>
</tr>
<tr>
<td>APPLI</td>
<td>---</td>
<td>---</td>
<td>---</td>
<td>---</td>
<td>0.278</td>
<td>0.454</td>
<td>0.114</td>
</tr>
<tr>
<td>TESTI</td>
<td>---</td>
<td>---</td>
<td>---</td>
<td>---</td>
<td>---</td>
<td>0.485</td>
<td>0.010</td>
</tr>
<tr>
<td>ORGAN</td>
<td>---</td>
<td>---</td>
<td>---</td>
<td>---</td>
<td>---</td>
<td>---</td>
<td>---</td>
</tr>
<tr>
<td>GRADI</td>
<td>---</td>
<td>---</td>
<td>---</td>
<td>---</td>
<td>---</td>
<td>---</td>
<td>---</td>
</tr>
</tbody>
</table>

**MOTIV = ABILITY TO MOTIVATE**
**HANDL = HANDLING QUESTIONS AND CLASS DISCUSSION**
**DELIV = DELIVERY**
**APPLI = KNOWLEDGE AND APPLICATION OF SUBJECT MATERIAL**
**TESTI = TESTING**
**ORGAN = ORGANIZATION**
**GRADI = GRADING**
### Table 4

**Interrater Reliabilities of Harari-Zeck BARS and Modified BARS by Class**

<table>
<thead>
<tr>
<th>Class</th>
<th>Harari-Zedock Reliabilities</th>
<th>Spearman-Brown Reliabilities*</th>
<th>Modified BARS Reliabilities</th>
<th>N</th>
<th>Spearman-Brown Reliabilities*</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>.3462</td>
<td>.76</td>
<td>.5387</td>
<td>16</td>
<td>.95</td>
</tr>
<tr>
<td>B</td>
<td>.3497</td>
<td>.83</td>
<td>.4999</td>
<td>14</td>
<td>.93</td>
</tr>
<tr>
<td>C</td>
<td>.3464</td>
<td>.83</td>
<td>.3745</td>
<td>6</td>
<td>.78</td>
</tr>
<tr>
<td>D</td>
<td>.3052</td>
<td>.83</td>
<td>.3232</td>
<td>10</td>
<td>.83</td>
</tr>
<tr>
<td>E</td>
<td>.3668</td>
<td>.86</td>
<td>.4462</td>
<td>12</td>
<td>.91</td>
</tr>
<tr>
<td>F</td>
<td>.3316</td>
<td>.89</td>
<td>.4989</td>
<td>12</td>
<td>.92</td>
</tr>
<tr>
<td>G</td>
<td>.4647</td>
<td>.89</td>
<td>.4766</td>
<td>15</td>
<td>.93</td>
</tr>
</tbody>
</table>

*Yields reliabilities based on sample size (N)*
Table 5

Dimension Means and Standard Deviations Across Rates

<table>
<thead>
<tr>
<th>Dimension</th>
<th>MOTIV</th>
<th>DELIV</th>
<th>KNOWL</th>
<th>RELEV</th>
<th>RELAT</th>
<th>WORKL</th>
<th>TESTI</th>
<th>ORGAN</th>
<th>GRADI</th>
</tr>
</thead>
<tbody>
<tr>
<td>Harari-Zedeck BARS</td>
<td>4.60</td>
<td>4.83</td>
<td>5.54</td>
<td>5.05</td>
<td>5.99</td>
<td>5.03</td>
<td>4.82</td>
<td>5.24</td>
<td>5.44</td>
</tr>
<tr>
<td>Modified BARS</td>
<td>4.73</td>
<td>5.71</td>
<td>4.88</td>
<td>5.50</td>
<td>4.85</td>
<td>5.24</td>
<td>5.87</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

MOTIV = ABILITY TO MOTIVATE STUDENTS
DELIV = DELIVERY
KNOWL = DEPTH OF KNOWLEDGE
RELEV = RELEVANCE
RELAT = INTERPERSONAL RELATIONS WITH STUDENTS
WORKL = ASSIGNMENTS AND WORKLOAD
TESTI = TESTING
ORGAN = ORGANIZATION
GRADI = GRADING
HANDL = HANDLING QUESTIONS AND CLASS DISCUSSION
APPLI = KNOWLEDGE AND APPLICATION OF SUBJECT MATERIAL
REFERENCES


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


Silverman, S.B., & Wexley, K.N. Reaction of employees to performance appraisal interviews as a function of their participation in rating scale development. Personnel Psychology, 1984, 37, 703-710.


INSTRUCTIONS TO Raters

On the following pages you will find a number of scales designed to measure the teaching performance of a university-level psychology instructor. On each page you will find one aspect (or dimension) of teaching performance accompanied by a definition of that dimension. Below each dimension title and definition you will find a rating continuum with behaviors placed at various points beside it, indicating the judged desirability of each behavior. These examples of teaching behaviors and their corresponding scale values are only meant to give you a frame of reference for the desirability of certain behaviors. Your instructor may have never actually displayed some of these behaviors, but they are placed along the continuum to give you an idea of the scale value of commonly-observed teacher behaviors.

For each of the following dimensions of performance, you are being asked to think of 2 behaviors relevant to that dimension that your instructor might be expected to display based upon his/her behaviors in the past. These newly-written behaviors do not necessarily have to be actions which were actually observed; they need only be behaviors which are typical of your instructor's past behavior. After each of these newly-written behaviors, please put your own judgment of the scale value (on a 1-7 scale) of the behavior. You may write in the behaviors and their corresponding values anywhere on the page, since there may not be quite enough room along the continuum. If you have any questions concerning this procedure, please feel free to come forward and ask. Thank you very much.
ABILITY TO MOTIVATE STUDENT - the professor's generation of student's 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 psychology course with this professor, most students could be expected to enroll in other classes that deal with the field of psychology.

In this class, if a student hesitatingly 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!"

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 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.
DEPTH OF KNOWLEDGE— the professor’s mastery of the subject matter.

This professor 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.

This professor, when asked about where to look for material on a specific topic, 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 investigate.

This professor could be expected to handle a question that comes up during lecture that isn't covered satisfactorily in the assigned readings.

If a student happened to read an article about classical conditioning and asked his professor 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.

If a student asks this professor to elaborate on a specific theory, the professor could be expected to provide a general outline but must refer the student to other sources for specifics.

This professor 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.

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

This professor could not be expected to be able to add any original thoughts to the material that is available to the student in the text.

Instead of admitting that he does not know the answer to a question, this professor could be expected to offer a reply in vague, general terms that confuses the students even more.
RELEVANCE - the professor's relating of the subject matter to things important and meaningful to students.

When this professor lectures on social norms and role playing, he could often 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 became interested in discussing the psychological effects of power and authority in the Armed Forces.

This 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 when discussing "conformity" and "deviance 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 could be expected to try to avoid or dismiss specific questions relating to controversial issues.

This professor, even when specifically asked to relate some of his lectures on motivation to human beings instead of white rats, could be expected to ignore this request and not mention motivation of humans once during the semester.
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 the 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 class, students could be expected to have no difficulty understanding this professor's lecture on conditioned-response sets, but they could often be expected to be bewildered when he discusses 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 this professor's exams, students could be expected to go to the TA's 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.
ORGANIZATION - the professor's design of the course and his arrangement of the material, both in class and in out-of-class assignments.

This professor could be expected to integrate the reading material and out-of-class assignments with his lectures.

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.

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 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.
INTERPERSONAL RELATIONS WITH STUDENTS - the professor's rapport with and sensitivity to students.

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

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.

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

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.

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

If a student asks this professor to help him with class material 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.

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

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 location, students must ask him individually.

In this case, 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 "I-really-don't-care-if-you-do-it-or-not."

This professor could be expected to try to humiliate or embarrass students who disagree with him.
GRADING - the ways and purposes for which the professor uses grades.

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 not 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 to always 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 four-page papers on critical topics covered during the course.

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 test and an "A" on the final exam, the professor could be expected to give the student an "A" for the course because the student's average grade is less than an "A."

Each student in this 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.
ASSIGNMENTS AND WORKLOAD- quantity and quality of the class requirements.

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

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.

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

This professor could be expected to require three hours of problem-solving activity per week, but two hours of this consists of section meetings with TA's helping and only one hour of homework.

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.

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.

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.

This professor could be expected to not only assign five hours' worth of work per week, 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 understood the previous chapters.
TESTING - the ways and purposes for which the professor uses tests.

This professor could be expected to give a test on a recent topic such that the students often feel that they have learned something new about the topic just from taking the test.

If this professor gives an exam on psychological theories, 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 questions and half essay.

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

This 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.
Appendix B

INSTRUCTIONS TO RATERS

On the following pages you will find a number of scales designed to measure the teaching performance of a university-level psychology instructor. On each page you will find one aspect (or dimension) of teaching performance accompanied by a definition of that dimension. Below each dimension title and definition you will find a rating continuum with behaviors placed at various points beside it, indicating the judged desirability of each behavior. These examples of teaching behaviors and their corresponding scale values are only meant to give you a frame of reference for the desirability of certain behaviors. Your instructor may have never actually displayed some of these behaviors, but they are placed along the continuum to give you an idea of the scale value of commonly-observed teacher behaviors.

For each of the following dimensions of performance, you are being asked to think of 2 behaviors relevant to that dimension that your instructor might be expected to display based upon his/her behaviors in the past. These newly-written behaviors don't necessarily have to be actions which were actually observed; they need only be behaviors which are typical of your instructor's past behavior. After each of these newly-written behaviors, please put your own judgement of the scale value (on a 1-7 scale) of the behavior. You may write in the behaviors and their corresponding values anywhere on the page, since there may not be quite enough room along the continuum. If you have any questions concerning this procedure, please feel free to come forward and ask. Thank you very much.
TESTING - The professor's ability to design exams which test more than the student's ability to memorize the material.

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

- This professor could be expected to give multiple choice exams which require students to recognize verbatim statements of the assigned chapters.

- The students who perform best on this professor's tests could be expected to be those who can memorize the material the most.
ORGANIZATION- The professor’s design of the course and his arrangement of the material, both in class and in out-of-class assignments.

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.

When this professor’s 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 re-running and correcting errors in input.
GRADING - The professor's grading policies concerning examinations and assignments turned in late.

This professor could be expected to not mark off for papers which are a few days late; and when grading an exam, if the student is near the "borderline," he could always be expected to receive a higher grade.

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.

Each student in this 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.
ABILITY TO MOTIVATE STUDENT- The professor's generation of students' interest in the subject material.

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

In this 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!"

This professor could be expected to pose questions and issues to students that are later discussed in section meetings or with classmates and friends after class.

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.
HANDLING QUESTIONS AND CLASS DISCUSSION - The professor's skill at answering students' questions and promoting open discussion.

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

This professor 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.

This professor could be expected to try to avoid or dismiss specific questions relating to controversial issues.

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

This professor could be expected to try to humiliate or embarrass students who disagree with him.
DELIVERY- The professor's ability to properly pace his lectures and present them in an understandable way.

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

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

In order to study for this professor's exams, students could be expected to go to the TA's because they can't understand the explanations of the professor.
KNOWLEDGE AND APPLICATION OF SUBJECT MATERIAL - The professor's mastery of the subject material and ability to relate this knowledge to things important and meaningful to the students.

When this professor 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.

If a student asks this professor to elaborate on a specific theory, this professor could be expected to provide a general outline, but must refer the student to other sources for specifics.

This professor could not be expected to be able to add any original thoughts to the material available in the text.

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