Effects of Verbally Labeled Anchor Points on the Distributional Parameters of Rating Measures

Grace French-Lazovik and Curtis L. Gibson
University of Pittsburgh

The hypothesis was examined that the negative skew found in most distributions of performance ratings is a function of the verbal labels used as anchors. When verbal labels quantified on the basis of the range of real-life performance were employed, distributional parameters (means, skewness) were affected. Typically used sets of labels were shown to be more negative than believed, thus tending to force responses toward the high end of the scale and thereby contributing to negative skew.

The use of rating scales to obtain measures of performance has substantially increased in recent decades in basic research investigations and especially in applied evaluation research. Despite the large amount of research on rating methodology, some of the measurement parameters of these scales are not well understood. (Exhaustive reviews of this literature are provided by Wherry, 1950, for research published prior to 1950, and by Landy & Farr, 1980, for studies published between 1950 and 1978.) In particular, the factors that might bring some degree of control over the distributional parameters of rating scales have received little attention.

Typically, rating scales provide a specified number of categories, usually representing a continuum, so that a respondent may choose one of the categories to express a judgment about some characteristic of an object or of human behavior. The term "rating scale" is generic and refers to a variety of methodologies for obtaining judgments, including the Semantic Differential (Osgood, 1952), the Q-sort (Stephenson, 1953), and the commonly used numerical rating scales. The principal ways in which these methods differ are in the number of anchor points (or defined categories) along the scale and the way in which the anchor points are described. Some of the most frequently used descriptors are verbal labels, marked lines between endpoints, integers, or integers in conjunction with verbal descriptors. It is commonly believed that numerical rating scales employing both integers and verbal descriptors provide the best aid to raters in defining scale positions.

Historically, with almost all rating scales, negatively skewed distributions of obtained ratings have resulted when human beings judge the overt behaviors of other humans in evaluative terms. The principal explanation in psychometric literature of this negative skew is the so-called "error of leniency" (Guilford, 1936), which refers to an assumed tendency on the part of a rater to be lenient in his or her ratings of other people. Postulating that other factors may contribute to the negative skew phenomenon, this research investigated the hypothesis that *the distributional shape obtained*

APPLIED PSYCHOLOGICAL MEASUREMENT
Vol. 8, No. 1, Winter 1984, pp. 49–57

0146-6216/84/010049-09$1.70

from rating scales is dependent, in part, upon the choice of verbal labels used to anchor rating scale points.

A very commonly used evaluative rating scale in the behavioral sciences, and in industry as well, employs the following set of verbal anchorings: Very _____, Above Average, Average, Below Average, Very _____. Even those formats that differ from this one often have the label "average" as the midpoint of the scale. However, when ratings of human behaviors are made, the term "average" may be perjorative; people typically do not like to be described by this term. Thus, average may not be a true midpoint on a scale. Using a verbal label that is not a midpoint to anchor the middle category essentially forces the higher categories to be used more frequently than the middle. Consequently, the distribution of obtained scores is shifted higher on the scale, and a negatively skewed distribution is generated.

Investigating this problem necessitated bringing some objectivity to the choice of verbal descriptors. Other researchers have chosen labels by selecting words that have been psychophysically scaled as to the degree of favorableness they represent. An alternative approach, one more analogous to procedures used in test construction, would be the representation of a continuum based on the words that raters typically apply to the performance they are judging (in this instance by teachers) rather than on the continuum indicated by psychophysically scaled anchor words. An empirical procedure was developed to derive numerical values, called descriptor indices, that locate the position of each of a set of possible verbal labels on a continuum by determining the proportion of teachers that potential raters would characterize by each label. It is hypothesized that these descriptor indices would function in a manner somewhat similar to item difficulty levels in test construction. By choosing verbal labels to anchor scale points on the basis of descriptor indices, two different rating scales were constructed and the relationship between their distributional parameters investigated. Evaluations of teaching made by college students provided the testing ground for this research.

Study I—Descriptor Indices

Method

Instrument. A set of 37 verbal labels culled from student evaluation of teaching rating scales was chosen in such a way as to ensure that the entire range of the continuum was spanned and the words were adequately familiar to college students (Thorndike & Lorge, 1944). The instrument for obtaining data necessary to compute descriptor indices consisted of instructions, four examples of the task, and two sections presenting the same verbal labels, in two different random orders, to which subjects responded. In Section I each subject indicated the percentage of instructors in his or her high school or college career that he or she would characterize as falling (1) below and (2) at or above each of the verbal labels. Section II required subjects to respond, not from their own experience, but as they believed a *typical student* would respond. Section II was included to draw students' attention to the fact that their individual experiences might differ from those of a typical student. In particular, the term "average" was expected to be troublesome in that subjects might automatically indicate 50% falling above and below this point. Forcing subjects to attend to their own experiences was an attempt to hold such an effect to a minimum.

Subjects. Responses were obtained from 121 students taking courses in introductory psychology, statistics for psychology, and abnormal psychology.

Analyses. The percent indicated as falling below each label was pooled across subjects, and the mean percent and standard deviation were calculated separately for Sections I and II. On the basis of the mean percentages, verbal labels were rank ordered for each section and a Spearman rho computed to determine similarities in the two orderings. The standard deviation for each label was used as an indicator of the degree of agreement among respondents as to its scale position; the mean percent falling below each label in Section I was designated its "descriptor index," and these indices

were related to integer scale points by the following formula:

$$d = 100\left[1 + \frac{p - n}{n - 1}\right] \quad (1)$$

where d = descriptor index,
p = a scale point, and
n = number of scale points,
i.e., $p = 1, 2, 3, \ldots, n$.

Results

Table 1 presents the summary statistics on the scale positions of 37 verbal labels calculated separately for the two different instructional sets, i.e., from the individual's experience and the typical student's view. Very high agreement between the two orderings of labels is demonstrated by the high Spearman rho ($r = .989$). From the *SD*s it can be seen that there is greater agreement among individuals about the positions on the continuum of very favorable or very unfavorable labels than for labels near the middle of the scale. However, agreement is greater for the label "average" than for most of the other labels near that point. A t statistic, calculated to test whether the descriptor index for "average" was significantly different from the theoretical midpoint of 50% gave the value $t(107) = -4.749$, $p < .005$. Thus, the descriptor index for the label "average" falls below the scale midpoint, as hypothesized.

Attention should also be drawn to the labels "above average" and "below average." In terms of the descriptor indices, the distance from "above average" to "average" is 21.77%, which is approximately equal to the distance from "average" to "below average" (18.74%). This set of verbal labels characterizes a scale whose points are roughly equidistant, but the entire set is displaced toward the unfavorable end of the continuum, that is, not centered about the midpoint.

Study II—Rating Scales

Method

Instruments. Two teacher rating questionnaires were constructed using identical item stems and instructions; they differed only in the labels attached to scale points. Five-point numerical scales on which each point was labeled with both an integer and a verbal descriptor were attached to a global rating item and to one item representing each of the three factors that predict the global one (French-Lazovik, 1974). The four items chosen were (1) Explanations were clear and understandable, (2) Material was presented in an interesting manner, (3) Student's intellectual curiosity was stimulated, and (4) Overall teaching ability.

Using the descriptor indices obtained on the basis of the first phase of data collection, labels were chosen for the two questionnaires, as illustrated in Table 2. It can be seen by comparing the descriptor indices (rounded to two digits) for the two sets of labels that those for the second questionnaire (Q_2) are approximately one scale point lower than those for the first questionnaire (Q_1). Thus, if descriptor indices do, in fact, act in a way similar to difficulty levels in test construction, then the mean numerical rating on Q_1 should be larger than on Q_2 for each item rated.

Subjects. Ratings on the four items were made by 304 students in the classes of six different instructors.

Procedure. Instructors were not present in their classes while the questionnaires were administered by a special proctor during the first 15 minutes of a class hour. Following instructions that guaranteed student anonymity, the two questionnaires were distributed in such a manner that half the students within a class randomly received Q_1, and half received Q_2.

Data analysis. The question, "Do the verbal labels used to anchor the scale points of rating scales influence the ratings given?" was answered through separate analyses of variance for each item stem, using a randomized block design (Table 3). Classes were used as blocks in order to control for between-instructor variability.

Results

Table 4 contains the paired t tests for assessing differences in the mean ratings resulting from the

Table 1
Summary Statistics of the Proportion of Instructors Characterized As Falling Below Each Verbal Label

	Section I				Section II			
Verbal Label	Mean %: Descriptor Index	SD	N	Rank Order	Mean %	SD	N	Rank Order
Exceptional	91.311	8.707	119	1	92.378	6.263	111	1
Superior	90.884	6.844	121	2	91.298	6.522	114	3
Outstanding	90.681	7.167	119	3	91.578	6.660	116	2
Excellent	90.432	8.675	118	4	90.848	7.180	112	4
Very High	83.786	10.564	117	5	86.947	11.215	114	5
Very Good	79.217	10.922	120	6	79.862	10.477	109	9
High	79.147	10.669	116	7	82.574	10.408	115	6
Very Competent	76.860	15.099	107	8	77.466	19.361	116	10
Very Effective	76.783	12.685	115	9	82.514	11.060	111	7
Well Above the Average	75.853	12.544	116	10	81.561	10.867	107	8
Better than Most	74.919	15.687	111	11	70.835	17.147	109	11
Above Average	66.923	14.884	117	12	67.512	16.896	121	12
Fine	61.407	19.189	108	13	56.061	19.985	115	15
Effective	57.875	15.748	104	14	56.588	20.033	114	14
Good	56.514	13.600	107	15	58.845	16.225	116	13
Moderately Good	54.520	15.166	102	16	52.692	17.676	120	16
Compares Well with the Average	53.705	15.587	105	17	51.233	17.085	120	18
Middling	47.374	10.413	115	18	45.809	11.161	115	23
Competent	47.000	20.714	119	19	48.488	20.013	121	20
Satisfactory	46.602	17.045	113	20	51.446	20.287	121	17
Needs Improvement	45.880	30.642	117	21	51.059	28.833	118	19
Average	45.157	10.606	108	22	47.383	11.341	115	21
Adequate	40.441	16.669	111	23	46.458	17.718	120	22
Mediocre	38.926	13.866	121	24	41.906	14.791	117	24
Only Fair	32.212	12.570	104	25	38.248	14.076	117	25
Below Average	26.419	12.519	114	26	30.791	13.303	110	26
Unsatisfactory	19.692	12.394	117	27	25.481	12.396	104	27
Ineffective	19.336	11.769	116	28	22.730	12.613	115	28
Weak	18.782	11.572	119	29	21.770	11.711	113	29
Low	18.308	10.389	120	30	16.113	10.578	115	33
Ranks Below Most	17.802	13.860	111	31	19.942	12.661	103	30
Inadequate	17.188	11.434	117	32	18.384	11.429	112	31
Poor	15.250	10.386	120	33	16.664	9.937	112	32
Inferior	14.529	9.651	119	34	14.179	9.315	112	36
Very Low	13.442	9.068	120	35	15.441	8.919	111	34
Very Ineffective	12.356	9.532	118	36	14.708	10.004	106	35
Very Poor	11.033	8.517	120	37	11.640	8.293	114	37

two questionnaires. Additionally, when the individual class means are compared (Table 5), Q_1 is larger than Q_2, as hypothesized, in 21 of the 24 possible comparisons (four items in each of six classes). The three deviations from this hypothesis all occurred in the same class, and the authors believe that an event that took place while the ratings were being made contributed to carelessness on the part of the students. (The instructor, waiting outside the classroom, posted grades from a recent exam; students hurriedly finished the ratings in order to go out and see their grades). Clearly, the mean ratings from Q_1 and Q_2 are different, and in the hypothesized direction.

Table 2
Verbal Labels and Their Descriptor Indices (DI) for Questionnaires 1 and 2

Integer Anchor	Q_1 Verbal Label	Q_1 DI	Q_2 Verbal Label	Q_2 DI
1	Exceptional	91	Very Effective	77
2	Very Good	79	Above Average	67
3	Good	57	Average	45
4	Satisfactory	47	Below Average	26
5	Unsatisfactory	20	Very Ineffective	12

Table 3
Randomized Block ANOVA of Differences in the Mean Ratings for the Two Teacher Rating Questionnaires

	Source of Variation	Sum of Squares	df	Mean Square	F
Stem 1.	Explanations were clear and understandable.				
	Class	89.276	5		
	Questionnaire	6.858	1	6.858	8.496**
	Residual	240.569	298	0.807	
	Total	336.557	304		
Stem 2.	Material was presented in an interesting way.				
	Class	139.367	5		
	Questionnaire	14.879	1	14.879	18.939**
	Residual	234.119	298	0.786	
	Total	387.948	304		
Stem 3.	Student's intellectual curiosity was stimulated.				
	Class	99.053	5		
	Questionnaire	8.297	1	8.297	9.550**
	Residual	258.917	298	0.869	
	Total	366.013	304		
Stem 4.	Overall teaching ability.				
	Class	87.980	5		
	Questionnaire	5.513	1	5.513	6.677**
	Residual	246.066	298	0.826	
	Total	339.390	304		

**$p < .01$.

In addition to the means, the reliabilities (Table 4), calculated by Horst's (1949) generalized formula, and two other distributional parameters, the variance and the coefficient of skewness (Table 5), were examined. Although the variances for the two questionnaires did not differ significantly, the coefficients of skewness are larger for each item stem for Q_2 than Q_1. Reliabilities were high for both scales.

Finally, all of these effects are visually represented for Item 4, overall teaching effectiveness, in Figure 1. (Plots of the other items give nearly identical results.)

Discussion

Clearly, the degree of negative skew in distributions of behavioral rating measures can be altered by the verbal labels used as anchors. Both means and skewness coefficients were affected by the verbal labels used in this research. Data presented also show that a rating scale anchored by a

Table 4
Item Reliabilities and Paired t-tests for Differences in the Mean Ratings on the Two Teacher Rating Questionnaires

Stem	$\bar{Q}_1 - \bar{Q}_2$	$SD_{(Q_1-Q_2)}$	t(5df)	Reliability Q_1	Reliability Q_2
1. Explanations were clear and understandable.	.310	.089	3.481**	.90	.88
2. Material was presented in an interesting way.	.458	.100	4.576**	.93	.92
3. Student's intellectual curiosity was stimulated.	.323	.062	5.165**	.88	.90
4. Overall teaching ability.	.273	.123	2.223*	.91	.87

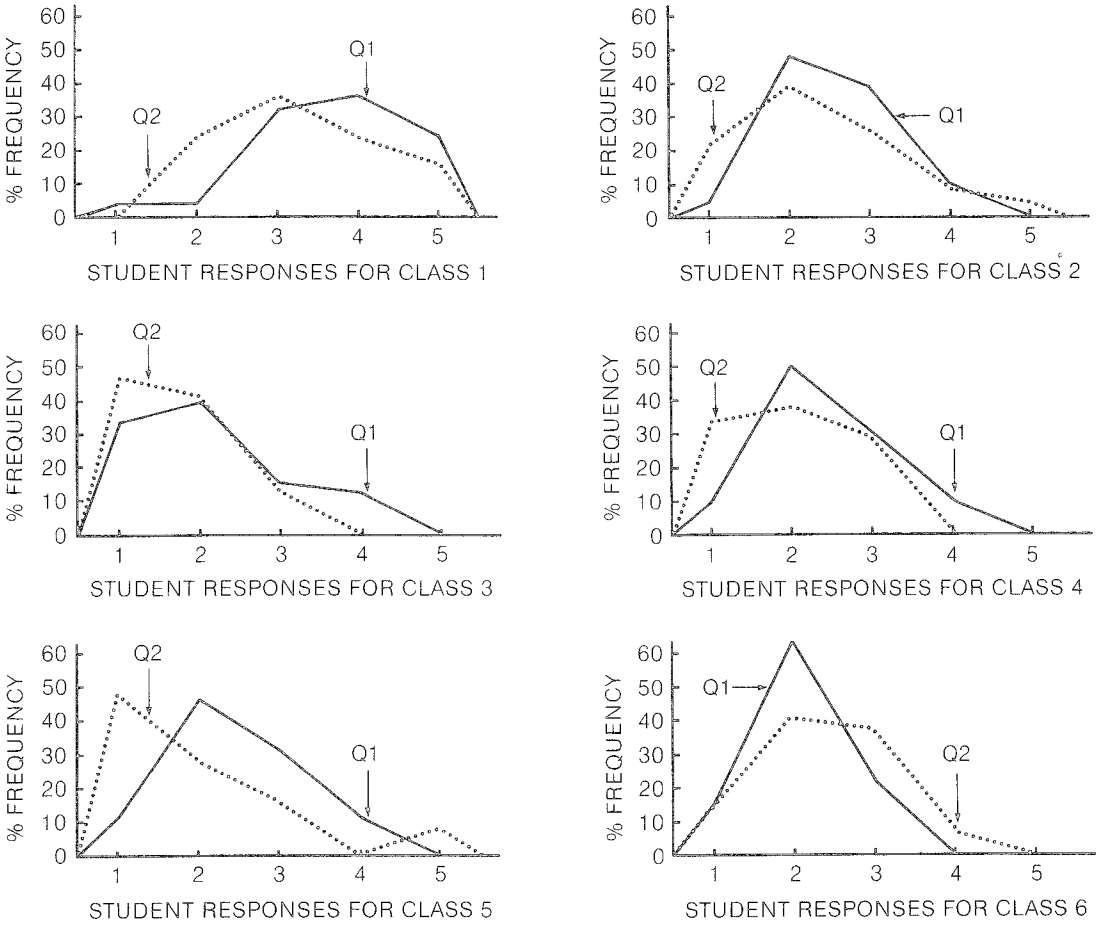
*p < .05. **p < .01.

Table 5
Means, Variances, F for Variances, Wilcoxon W, and Coefficients of Skewness and Kurtosis for Questionnaires 1 and 2

	N		Mean		Variance				Skewness		Kurtosis	
Class	Q_1	Q_2	Q_1	Q_2	Q_1	Q_2	F	W	Q_1	Q_2	Q_1	Q_2
Item 1:	Explanations were clear and understandable.											
1	25	25	3.560	3.120	1.090	1.277	1.172	1.556	-0.650	0.312	0.140	-0.670
2	21	23	2.571	2.304	0.957	0.676	1.416	0.748	0.311	-0.110	-0.986	-0.576
3	33	32	1.848	1.594	0.945	0.572	1.652	1.038	1.407	0.855	2.369	-0.673
4	20	21	2.000	1.571	0.632	0.657	1.040	1.974	0.699	1.613	0.807	2.821
5	26	25	2.462	1.920	0.658	0.910	1.383	2.088*	0.377	0.483	-0.206	-1.080
6	27	27	2.074	2.148	0.533	0.823	1.561	-0.253	0.525	0.356	0.733	-0.565
Item 2:	Material was presented in an interesting way.											
1	25	25	4.080	3.560	0.577	1.007	1.745	1.954	-0.759	-0.044	1.062	-0.971
2	21	23	2.762	2.130	0.890	0.664	1.340	2.136*	0.132	0.297	-1.227	-0.231
3	33	32	1.909	1.531	0.960	0.451	2.129*	1.537	1.040	0.903	0.302	-0.243
4	20	21	2.750	2.190	0.934	1.162	1.244	1.668*	0.559	0.378	0.176	-1.088
5	26	25	2.577	1.920	1.054	0.660	1.597	2.338**	0.016	0.660	-1.081	0.258
6	27	27	2.037	2.037	0.499	0.729	1.461	0.112	-0.052	0.326	-0.854	-0.636
Item 3:	Student's intellectual curiosity was stimulated.											
1	25	25	3.880	3.600	0.943	1.167	1.238	1.001	-1.226	-0.173	2.187	-1.181
2	21	23	2.619	2.261	0.648	0.747	1.153	1.539	-0.428	0.365	0.055	-0.219
3	33	32	2.303	1.875	1.218	0.500	2.436**	1.417	0.982	0.182	0.566	-0.890
4	20	21	2.650	2.476	1.082	0.762	1.420	0.485	0.491	0.329	-0.031	-0.409
5	26	25	2.308	1.760	0.622	1.273	2.047	2.923**	0.433	2.208	0.147	4.083
6	27	27	2.519	2.370	0.721	0.858	1.190	0.384	0.345	-0.217	-0.463	-0.923
Item 4:	Overall teaching ability.											
1	25	25	3.720	3.320	1.043	1.060	1.016	1.504	-0.655	0.280	0.607	-0.964
2	21	23	2.524	2.348	0.562	1.146	2.039	0.850	0.305	0.678	-0.075	0.276
3	33	32	2.061	1.656	0.966	0.491	1.967	1.576	0.676	0.600	-0.477	-0.714
4	20	21	2.400	1.952	0.674	0.648	1.040	1.585	0.355	0.090	-0.065	-1.417
5	26	25	2.423	1.920	0.734	1.410	1.921	2.291*	0.258	1.464	-0.362	1.809
6	27	27	2.074	2.370	0.379	0.704	1.858	-1.423	-0.036	0.021	-0.094	-0.445

*p < .05. **p < .01.

Figure 1
Comparison of Percentage Frequency Polygons for Questionnaires 1 and 2 on the Overall Teaching Effectiveness Item



set of verbal labels *more positive* than those typically used (having higher descriptor indices) results in a shift of the mean numerical rating toward the *less positive* end of the scale; conversely, rating scales that are anchored by a set of less favorable labels result in mean numerical values shifted toward the favorable end of the scale. A finding of Lam and Klockars (1982) bears some similarity to this finding. Using entirely different methodology and investigating a different question about rating scales, their results indicate that the relationship between scales having all intermediate points labeled verbally and those having only the endpoints labeled is a function of the psychophysical scale values of the intermediate labels. Among the rating scales they used, lower means resulted for scales whose response categories were described by labels with more favorable psychophysical scaling values; ratings with verbal labels of less favorable scaling values produced higher means.

The relationship of the observed nonsignificant differences in variability for Q_1 and Q_2 to the leniency problem is not a simple one. To understand this, it must be remembered that there are two

levels of variability in these ratings: first, there is variability among the students in a single class in their judgments of the teacher (i.e., $s^2_{\text{among judges}}$) and second, there is variability in the class means across teachers (i.e., $s^2_{\text{teacher means}}$). In order for ratings to have high reliability, the among-judges variability should be low, and the variance of mean ratings should be high, indicating ability to discriminate among those rated. That among-judges variability was not significantly different for Q_1 and Q_2 suggests that altering the verbal labels, at least for those in this research, did not result in greater disagreement among students within the same class in judging their teacher.

The number of teachers rated in this study is not large enough to address variability of mean ratings across teachers. However, the finding that mean ratings were affected by the verbal labels used strongly suggests that the failure to discriminate among various levels of good performance, which characterizes the leniency effect, would be lessened if labels were very carefully chosen.

Several findings that have implications for the construction of rating measures emerged from the calculation of descriptor indices. First, the use of the term "average" as a midpoint anchor is called into question, as well as its frequently associated labels "above average" and "below average." Although more research is needed with ratings of other complex behaviors, students judging teaching behaviors characterize significantly less than 50% of their teachers by the term "average."

Second, an inspection of the descriptor indices reveals an important result—that adjectives whose meanings are logical opposites do not characterize numerically opposite proportions of a scale. Compare, for instance, the descriptor indices for a scale that uses what appear to be logically balanced adjectives, such as Very Effective (DI = 77), Effective (58), Ineffective (19), and Very Ineffective (12). Although students characterize 23% of their teachers as falling at or above the label Very Effective, only 12% are characterized as falling below Very Ineffective. Similarly, 42% fall at or above Effective, only 19% at or below Ineffective. The same holds for other logically balanced descriptors, e.g., Very Good (79), Good (56), Poor (15), and Very Poor (11). The obvious lack of symmetry in the scale positions of these labels indicates that they are not at all the balanced sets they have been assumed to be if the scales are to represent the empirically determined range of performance. If the finding of adjective asymmetry extends to the judgment of other domains of human performance as well, then substantial questions arise about the results that have been obtained with scales employing such logically balanced anchors, including the use of polar adjectives to define scale endpoints.

From the point of view taken here, rating scale construction should begin by empirically determining the real range of performance in the behavior to be judged. This range can be ascertained by calculating a descriptor index for each of a set of possible verbal labels, just as test construction begins with the determination of item difficulty levels. Choice of labels on the basis of the obtained descriptor indices can then make possible the construction of scales that produce finer discriminations by raters within the portion of the response continuum of concern to the investigator.

These findings also have consequences for the interpretation of rating measures. Most often, mean ratings are compared to the numerical midpoint of their scale. If some of the common anchor descriptors have been used, then such a comparison could lead to severely distorted interpretation.

Although this research has shown some of the effects of verbal anchors on numerical ratings, there is no intention of implying that labels are the only determiners of rating distributions. The number of scale points, the number of verbal anchors, and the numerical values associated with each of the scale points are factors that may exert an influence on the distribution in some fashion.

In addition to casting doubt upon the "leniency effect" as the major determinant of negatively skewed rating distributions, these research findings suggest that whether rating measures represent the continuum of performance that occurs in the real world (as do achievement tests) or a continuum defined by logically balanced adjectives has non-

trivial psychometric consequences that need further examination in other domains of human performance.

References

French-Lazovik, G. (1974). Predictability of students' evaluations of college teachers from component ratings. *Journal of Educational Psychology, 66,* 373–385.

Guilford, J. P. (1936). *Psychometric methods*. New York: McGraw-Hill.

Horst, P. (1949). A generalized expression for the reliability of measures. *Psychometrika, 14,* 21–32.

Lam, T. C. M., & Klockars, A. J. (1982). Anchor point effects on the equivalence of questionnaire items. *Journal of Educational Measurements, 19,* 317–322.

Landy, F. J., & Farr, J. L. (1980). Performance rating. *Psychological Bulletin, 87,* 72–107.

Osgood, C. E. (1952). The nature and measurement of meaning. *Psychological Bulletin, 49,* 197–237.

Stephenson, W. (1953). *The study of behavior: Q-technique and its methodology*. Chicago: University of Chicago Press.

Thorndike, E. L., & Lorge, I. (1944). *The teacher's word book of 30,000 words*. New York: Teachers College, Columbia University.

Wherry, R. J. (1950). *Control of bias in rating: Survey of the literature* (Tech. Rep. No. DA-49-0853 OSA 69). Washington DC: Department of the Army, Personnel Research Section.

Acknowledgments

The authors thank Drs. Lloyd Bond, Charles A. Perfetti, and Mary A. Hartz for their critical reviews and valuable comments on this research. Special appreciation is expressed to Dr. Sara Strouss and Marian Skoog for help in preparing the manuscript and to Stephen Hennings for the preparation of Figure 1.

Author's Address

Send requests for reprints or further information to Grace French-Lazovik, Director, Office for the Evaluation of Teaching, 3600 Cathedral of Learning, University of Pittsburgh, Pittsburgh PA 15260, U.S.A.