THE USE OF RATIO PRODUCTION SCALES TO ASSESS QUALITY OF TEACHING PERFORMANCE[1]

Fred C. Feitler
and
Stephen A. Graf
Youngstown State University

[1] Prepared for presentation at the 1978 Convention of the American Educational Research Association, Toronto, Canada, March 27 - 31.

Introduction

This study attempts to show the feasibility of an alternative method to the frequently used category scaling procedure, when what is being judged are subjective dimensions such as student reaction to instruction. Category scaling is compared to ratio production (multiply-divide) scaling, with focus on the shortcomings of the former and the advantages of the latter.

Two critical questions are addressed involving two kinds of measurement scales that can be used in educational evaluation: 1) How appropriate is the scale from a scaling perspective? and 2) How appropriate is the scale from a user's perspective? A third question of lesser importance is also considered, i.e., to what extent do the different scales produce the same conclusion? The first question deals with the issue: What scaling assumptions are met, or are not met, by the use of this scale for this purpose? The second question deals with the issue: To what extent do individuals who use this scale for this purpose understand, or misunderstand, the results obtained from use of the scale? One should note that the validity of the conclusion, by whatever the scale, is not the issue. For example, appropriate use of a scaling technique might lead one to conclude that students perceive a particular instructor to be, on the average, two times less effective than other instructors at a university. Perceived effectiveness, however, might not coincide with other measures of instructor effectiveness, e.g., student achievement. Stated another way, it is entirely possible for organisms to reliably misperceive.

An example illustrates the nature of the problem. A faculty evaluation form item typically calls for students to make perceptual judgments about the quality of effectiveness of teaching they have received.

One such item might appear as:

> *The instructor's effectiveness in teaching the subject matter was:*
>
> *A = Excellent (Best 10%)*
> *B = Good (Next 20%)* *D = Poor (Next 20%)*
> *C = Average (Mid 40%)* *E = Very poor (Worst 10%)*

This category scale is arbitrarily weighted with A = 5, B = 4, C = 3, D = 2, and E = 1. All responses necessarily fall within the five categories, and the total within each category can be counted for any class and then combined for all classes. The number of responses in each category is multiplied by the weight assigned to that category. Then these weighted category products are summed and divided by the total number of individuals responding. An arithmetic mean is thus obtained for any instructor, any set of instructors, or all instructors.

Now, instead of simple counts of the number of responses in each category, a uni-measure, the arithmetic mean, is generated, which expands the original five categories into any number of possibilities between 1 and 5, depending on the number of decimal places. As a measure of central tendency, this one number describes an obtained dispersion of responses across the five categories. Several problems exist, however. The arbitrary assignment of weights forces an equal interval between all successive categories, independent of the description of the percentage distribution, i.e., 10%, 20%, 40%, 20%, 10%. The equal interval assumption says, for example, that the distance between the lowest "C" and the lowest "D" is exactly the same as the distance between the lowest "A" and the lowest "B". Such an assumption seems difficult to justify.

Another problem is the skewness typically found when one plots all of the obtained arithmetic means. For those who have engaged in Faculty Evaluation practice and/or research, this highly skewed curve is familiar. The interpretation of such a curve can be confusing. (See Figures 1 and 2) For example, it is not unusual to have a population mean of 3.9 (where 5.0 would be the highest possible mean), with a standard deviation of plus or minus one (approximately). This suggests that on the sample item, "good", instructor effectiveness is "average". Likewise, "average" instructor effectiveness is, relative to the arithmetic mean, "poor". This skewed distribution appears in a broad spectrum of research involving personal judgments of one person by another.

The use of standard scores or percentiles to describe an instructor's position further confuses interpretation. A 0.1 difference in an obtained arithmetic mean in the range of 1.5 - 2.5 may produce a one percentile point difference, while the same 0.1 difference in the range of 3.5 - 4.5 produces a ten percentile point difference. Also, an obtained arithmetic mean of 4.0, which is above the mean of the population (3.9), could place an individual in the 45th percentile.

To a faculty member at an institution where student reaction to instruction is part of the basis for granting tenure, deciding promotions, or making other crucial judgments concerning an individual's professional career, the questions raised by this type of category scaling treatment are not trivial.

From a scaling perspective, the skewness obtained by use of the category scaling technique suggests that the rated dimension, perception of instructor effectiveness in the example, is not normally distributed in the population. Thus some problems exist in use, understanding, and analysis of student ratings of instruction through category scaling.

Theoretical Framework

Ogden Lindsley, since 1965, has collected evidence (e.g., Lindsley, 1971) that behavior grows, changes, and spreads by multiplying. His development and refinement, with others, of the Standard Behavior Chart (e.g., Pennypacker, Koenig, and Lindsley, 1972) has demonstrated the utility of overcoming an add-subtract framework in looking at the world. On an add-subtract scale, equal

TABLE I. Category and Ratio Scale Distributions of All Responses and Class Means, By Standard Deviation Cutoffs, To Perceived Effectiveness of Instructor(s)

Type of Scale	Type of Distribution	Number of Cases Between Lower and Upper Standard Deviation Units -4,-3	-3,-2	-2,-1	-1,Mean	Mean,1	1,2	2,3	3,4
Category	All Responses	(impossible)	33 (truncated)	62	105	239	192 (truncated)	(impossible)	(impossible)
Category	Class Means	0 (truncated)	0	3	7	13	3 (truncated)	(impossible)	(impossible)
Ratio	All Responses	4	11	29	261	238	37	7	11
Ratio	Class Means	0	1	3	8	10	4	0	0

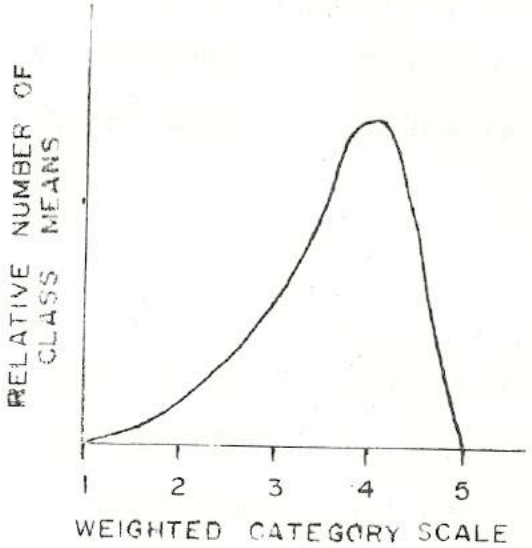

Figure I. Hypothetical category scale distribution of class means to a "perception of quality" item.

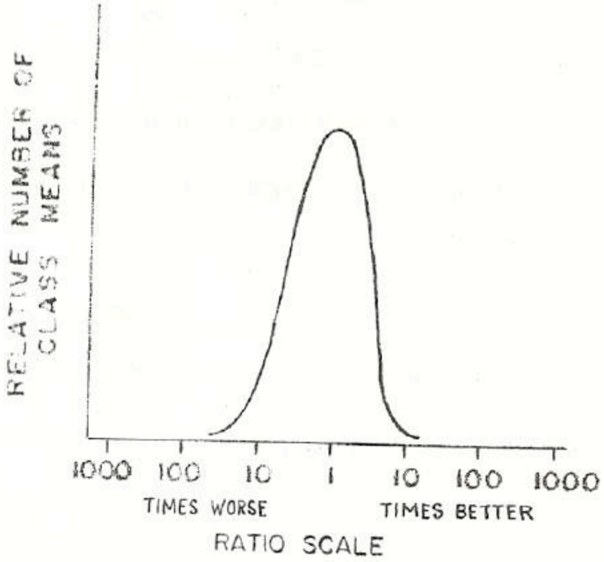

Figure 2. Hypothetical ratio scale distribution of class means to a "perception of quality" item.

distances represent equal differences. The distance between four and eight equals the distance between 100 and 104, and in each case, that distance, or difference is four.

On a multiply-divide scale, equal distances represent equal ratios. The distance between four and eight equals the distance between 100 and 200, and in each case that distance, or ratio, is two. The middle of the multiply-divide scale is one, with multiplying representing one direction away from one, and dividing representing the other direction away from one. For an add-subtract scale, zero is the middle, with positive numbers extending in one direction and negative numbers extending in the other direction. In both add-subtract and multiply-divide scales, the scale is symmetrical around the midpoint, zero and one, respectively. In other words, the distance from zero to +16 on an add-subtract scale is equal to the distance from zero to -16, and the distance from x1 to x16 on a multiply-divide scale is equal to the distance from ÷1 to ÷16. One should note that the middle of the multiply-divide scale can be represented by either x1 or ÷1. Multiplying any quantity by one does not change the quantity, and dividing any quantity by one does not change the quantity.

Having become familiar with Lindsley's discoveries of the multiplicative relationships replete in human behavior, and having involvement in a category scaling of instruction mandated for full-time faculty at the University, the authors came to examine a multiply-divide approach to the measurement of student perceptions of teaching effectiveness. The stimulus is represented by the various dimensions of instruction to which the student is exposed, while the response is the subjective judgment on the part of the student in reacting to the stimulus dimension. Such is the province of "psychophysics", named and founded by Gustav Fechner over 100 years ago. One of the foci of this area of perceptual psychology has been to try to show that there is a general psychophysical law relating subjective magnitude to stimulus magnitude. In the last four decades, S.S. Stevens has had perhaps the greatest impact on this field.

Stevens has contended (e.g., Stevens, 1957) that equal stimulus ratios produce equal subjective ratios, and that furthermore, subjects can make valid quantitative, or "direct" estimates of subjective events. Stevens' research involved perceptual continua such as loudness, brightness, taste, duration, heaviness, and visual distance, visual length, visual area, etc. Others (e.g. Ekman, 1956; Finnie & Luce, 1960) extended such findings to dimensions such as preference for various occupations, preferences for wrist watches, esthetic value of handwriting, drawings, music, seriousness of criminal offenses, etc. A common feature of this latter type of dimension is that the stimuli are measured on a nominal scale, rather than some type of metric scale. Common to both metric and nonmetric stimuli, however, is the feature that variability tends to increase in proportion to the apparent magnitude (Stevens, 1966). Lindsley would state this in more comprehensible terms, namely, judgments spread equally around a middle on a multiply-divide scale. In other words, on a multiply-divide scale, the distance from the middle to the top of the spread is the same as the distance from the middle to the bottom of the spread. Plugging in some numbers to generate an example, a middle of 12 with a multiply-divide spread of three would mean that the spread extends from 36 (12x3) to four (12÷3).

Following these lines of investigation, research was initiated in which, concurrent with the traditional category scaling techniques, students were asked to make judgments of instructional dimensions using multiply-divide ratios.

Method

A "Student Reaction to Instruction" questionnaire was administered by the Office of Instructional Improvement at Youngstown State University through its regular system at the end of Fall Quarter, 1976. The format utilized category scaling in which five-choice responses were described for each item. Student volunteers in each class were in charge and instructors were not present. Twenty-six experimental "ratio production (multiply-divide)" evaluations were administered in selected classes along with the regular form. The sample contained classes from 23 instructors selected randomly and three classes from the authors. Thirteen items were set up on the ratio production form to coincide with items on the regular form and an additional item asked for a judgment comparing the two forms. Two of the thirteen items were chosen for analysis because of their general applicability to all classes. The ratio production form required students to judge whether the particular dimension was more than, exactly the same as, or less than the specified standard, and then to fill in an appropriate number to represent the perceived ratio, in either the "more than" or "less than" case. As the instructions specified:

> *Note that the number you fill in is not restricted except that it must be greater than one. "One times greater" and "one times less" are identical to "exactly the same", since multiplying anything by one doesn't change it.*

"Zero" was explained to be an impossible response.

The instructor's effectiveness in teaching
the subject matter was:........................ a. _____ times more effective

b. □ exactly the same

c. _____ times less effective

than others I have had at this University.

Results

For all responses (over 600 students) on the category scale questionnaire to the statement "The instructor's effectiveness in teaching the subject matter was:", the arithmetic mean was 3.8 and the standard deviation was 1.1. The distribution of these responses appears in Fig. 3 by category grouping, and in Table 1 by standard deviation cutoffs. The arithmetic mean of instructor means (26 classes) was also 3.8. The standard deviation of these instructor means

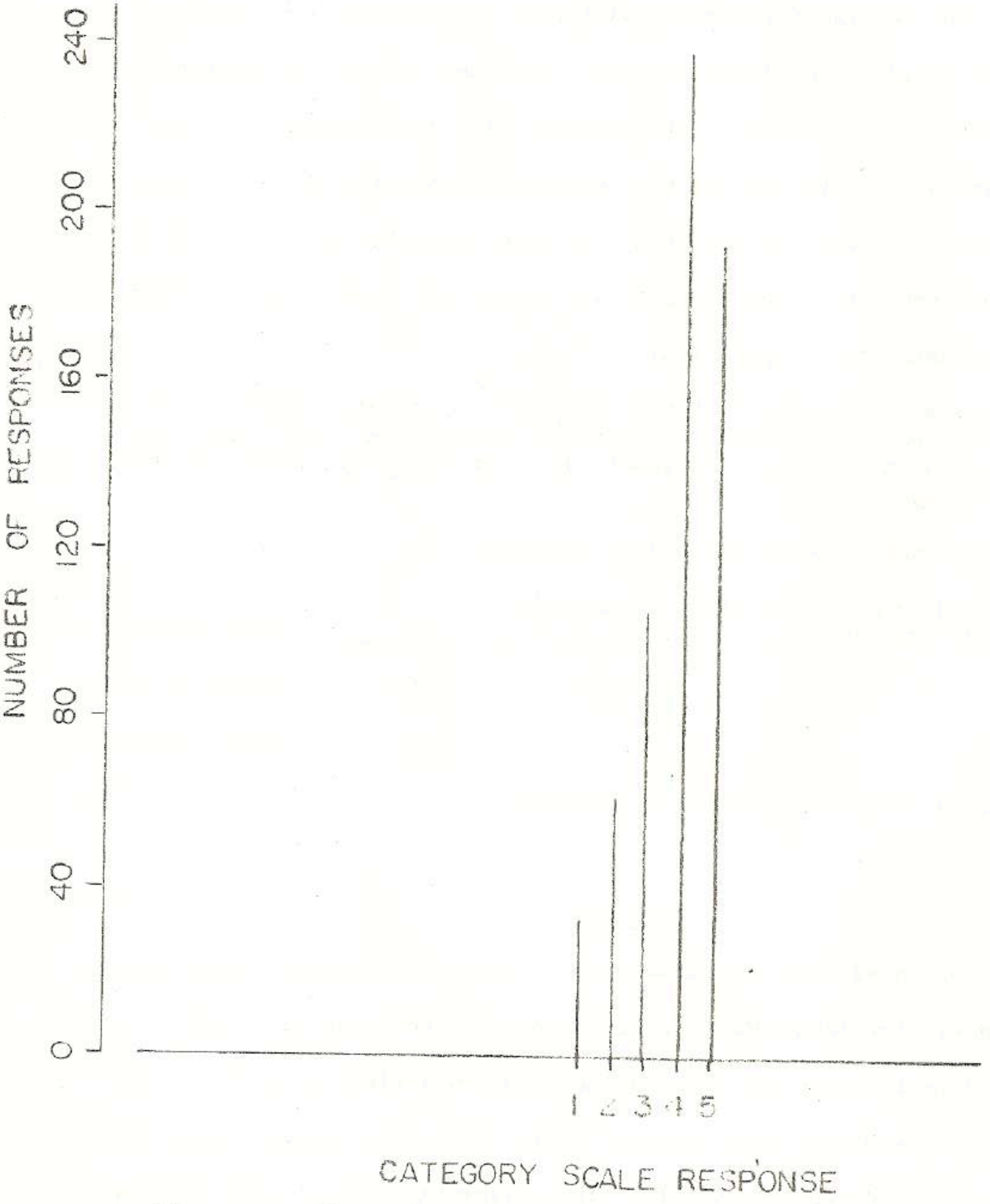

Figure 3. Category scale distribution of all responses to perceived effectiveness of instructors.

was 0.9. The distribution of these means appears in Table 1.

For all responses (about 600 students) on the ratio scale questionnaire to the statement: "The instructor's effectiveness in teaching the subject matter was:", the geometric mean was "1.4 times more effective", and the standard deviation was 4.2. The distribution of these responses appears in Fig. 4 in raw ratio form, and in Table 1 grouped by standard deviation cutoffs. The geometric mean of instructor means (26 classes) for the ratio scale was "1.5 times more effective", and the standard deviation was 2.2. The distribution of these means appears in Table 1.

A second statement dealt with perceived comparison of one's instructor to all other instructors had at the same university. The results obtained were very similar to those reported above. The Spearman rank-order correlation coefficient between an instructor's rank on perceived effectiveness and the same instructor's rank on perceived comparison was .96 using category scale ranks and .98 using ratio scale ranks for the two items. An instructor's rank on perceived effectiveness via the ratio scale correlated .85 with the same instructor's rank on perceived effectiveness via the category scale. An instructor's rank on perceived comparison to other instructors via the ratio scale correlated .90 with the same instructor's rank on the category scale.

The arithmetic mean for the category scale responses was obtained by multiplying the category weights (1,2,3,4, and 5) by the number of responses to each category in the overall sample (600 students), summing the products, and dividing by the number of students responding. As noted, the weights were arbitrary and chosen when the questionnaire was adopted, not by the students at the evaluation session. The standard deviation of the arithmetic mean for all responses was based on 631 students responding.

The arithmetic mean of instructor means was obtained by summing the mean of each of 26 classes and dividing by 26. The standard deviation of instructor means was based on the 26 class means.

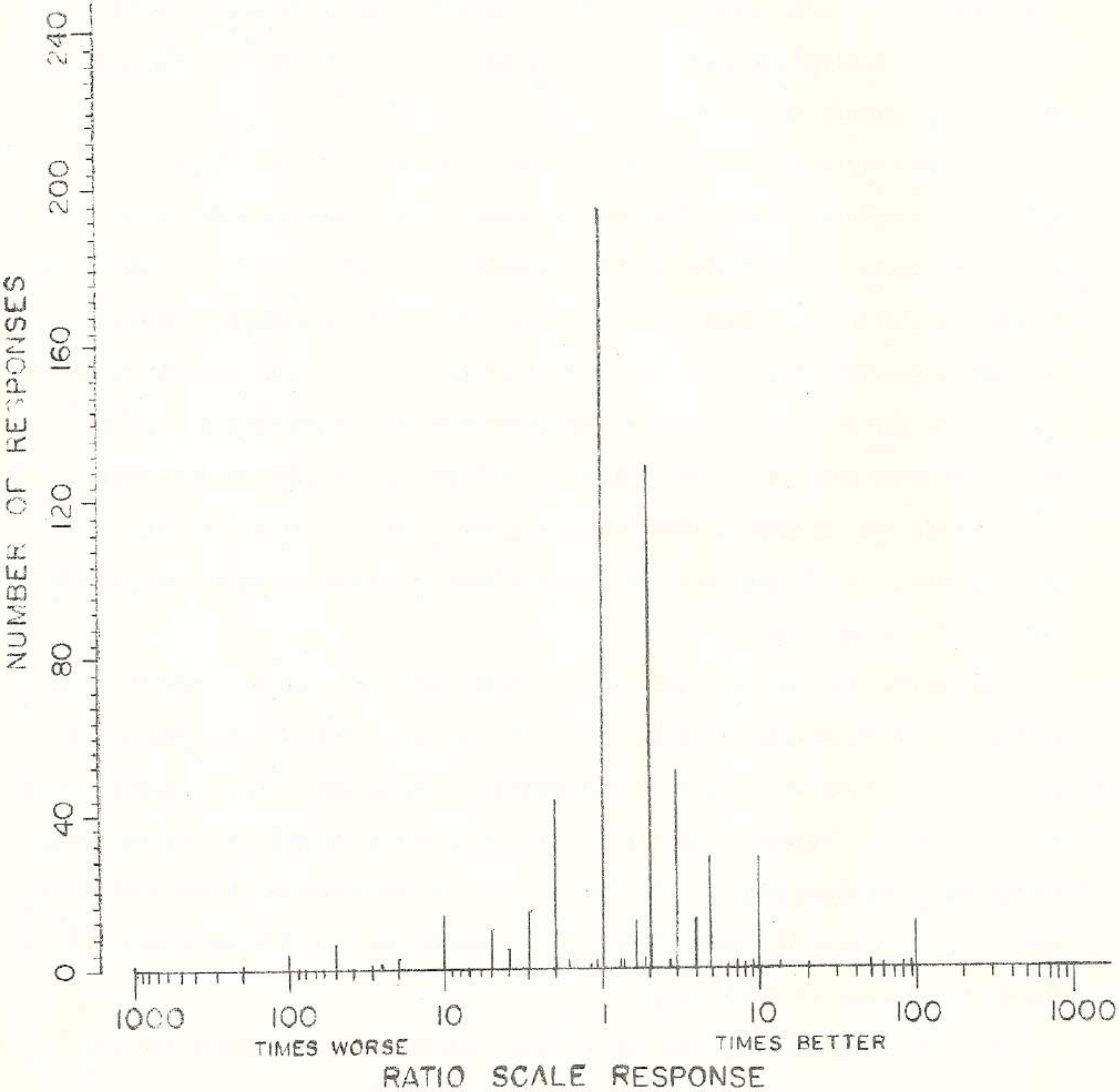

Figure 4. Ratio scale distribution of all responses to perceived effectiveness of instructors.

The geometric mean for all responses was obtained by transforming each student's ratio response (rounded to two significant digits) to a logarithm, summing the logs, dividing by the number of students responding, and transforming to an anti-log. For "divide by" responses, the ratio response was transformed to its inverse before being transformed to a log. A more direct method of obtaining the geometric mean involves multiplying (in the case of "times more than") and dividing (in the case of "times less than") each successive ratio to arrive at a total product, and then taking the _n_th root of that product, where _n_ is the number of students responding. The standard deviation of the geometric mean for all responses was based on 598 students responding.

The standard deviation of a geometric mean is similar to the more familiar arithmetic standard deviation in that both are measures of variability which describe, e.g., the middle two-thirds of a distribution, or the middle 95% of a distribution. While one expects to find about 68% of the scores in a distribution between the mean and plus and minus one standard deviation units of an arithmetic mean, one expects to find 68% of the scores between the mean and "one times" or "one divide by" the standard deviation units of a geometric mean. For a distribution with an arithmetic mean of 10 and a standard deviation of five, the score corresponding to one standard deviation above the mean would be 15, while the score one standard deviation below the mean would be 5. For a distribution with a geometric mean of 10 and a standard deviation of five, the score corresponding to one standard deviation above the mean would be 50 (i.e., one x five x 10), while the score one standard deviation below the mean would be two.

The geometric mean of instructor means was obtained by summing the log of the geometric mean of each of the 26 classes, dividing by 26, and transforming the result to an anti-log. The standard deviation of the instructor geometric means was based on the 26 class geometric means.

Discussion

Confusing Interpretation: Category scaling of student perception of teaching quality produces confusion in interpretation. The scale itself has one meaning by design; the distribution produces an entirely different meaning based upon the reported perceptions and the statistics generated from these data.

The use of a multiply-divide scale, producing ratios eliminates this dilemma. If it is necessary to compare an individual mean with a mean of another or with the mean of the distribution a ratio is derived which has a single interpretation.

An illustration is provided:

> Instructor "A" has a multiply-divide mean of x1.8 (better than). Instructor "B" has a corresponding mean of x2.4. The ratio produced 2.4/1.8 yields a ratio, instructor "B" is perceived to be 1.33 times better than instructor "A" on the dimension being rated. If the "grand mean", the mean for all instructors, was x1.2 (for illustration), then instructor "A" is perceived to be 1.5 times better than "average" and instructor "B" is perceived to be 2.0 times better than the average instructor.

Skewness: Category scaling of perceptual judgments is seen to produce a highly skewed distribution of means. The distribution of category frequencies is similarly biased. In contrast, the multiply-divide distribution of perceptions is more normally spread around the midpoint of the ratio scale. A larger sample than that provided by 26 classes (600 students) would further support or refute this point.

The distribution of responses for the category scaled item included 68% of the scores above the "average" category and only 15% below the average category. The multiply-divide distribution showed 10% "worse than", 33% "equal to" and 49% "better than". The tendency to rate instruction high rather than low is still apparent. However, there is a greater tendency to rate an instructor as the "same as" than to put him in the "average" category.

When the distribution of class means is examined around the "grand mean", the mean for all scores, the advantage of the multiply-divide scaling is more apparent. Using category scaling 13 classes fell one standard deviation above the mean, with 7 below. Using the ratio scale, 10 were within one standard deviation above; 8 fell within one standard deviation below the mean.

At least for this sample of 26 classes, this discussion suggests that when classes are the unit of comparison, there is a much greater tendency to approach a normal distribution of classes with the ratio scaling than with the category scaled measurement.

Comparable Statistical Treatment: Although generally less familiar to the educational practioner or administrator, the geometric mean and geometric standard deviation are analoges to those computed for add-subtract scales. These calculations can be handled by hand, with calculator, or computer.

One other advantage of the multiply-divide scaling procedure is that there is less likely to be truncated intervals with a ratio scale. A total scale range of one million is represented on the ratio scale as compared to the scale range of 5 for the category scale.

We have found it useful to show results graphically, on semi-log scales. Although the use of log scales is common in the sciences and engineering, they are less familiar to educators. Their use, however, facilitates understanding. Use of log scales or log depiction of data is not inherently difficult. Many learn to manipulate logs while in high school and have little or no use for them thereafter.

Content Comparability: Since there is increasing use of category scale data in personnel decision-making, it is apparent that there is confidence in the values generated. To compare the category scale results with those produced by ratio scales, correlation coefficients were produced. For the item: "How many times better (or worse) is this instructor than others you have had at this University?" There was a .901 rank order correlation between the two scaling approaches. For the effectiveness item cited earlier, there was a .854 correlation between the scales. These unusually high values suggest direct comparability.

Decision Making: The use of category scales can produce results with two levels of interpretation: 1) What does the scale mean? and 2) What does the distribution of scores mean? Using add-subtract scales, it is difficult to resolve this apparent contradiction. Because the ratio production scales produce ratios which can themselves be used as ratios, they provide a logical method to eliminate the possible dual interpretation. The statement that an instructor was judged by students to be 1.5 times better, or more effective than others in the department or the university is clear and can be used by a committee or administrator without concern for a possible second face value interpretation.

Student Response: Students tended to show a strong preference for the more commonly used category scale form. In fact they reported that they preferred it two times better than the ratio form. Was this because students are accustomed to responding to category scales? We can only guess, and suggest that further research be done in this area. Practice, warmup, or greater explanation might have reduced negative student reactions to the ratio production form. Since the instructions for filling out the form came from students who were themselves unfamiliar with the methodology, there is a possible residual effect produced by the administration.

General Application: Although the research reported in this paper deals only with student perceptions of teaching quality and issues of measurement related to this one area, we have a strong conviction that the problems addressed, concerned with interpretation of category generated measurement, have much broader application than to the issue of faculty evaluation alone.

The highly skewed curve described earlier is common to category scales which are used to judge perceptions of one person by another. The use of ratio scales as an alternative to category scales should be explored and evaluated in instances where perceptual judgments are involved. This could be extremely important for personnel evaluation in schools and elsewhere. Research involving student perceptions of "this or that" is common in the literature. Preference testing, valuing, and assessing are other broad areas where perceptions are often measured with category scales.

Suggested Research: Several areas have been identified in earlier discussion. There is a need to replicate this research with a larger sample. In the area of faculty evaluation, the class, rather than the individual student response becomes the unit of comparison. Although 26 classes are representative of all colleges and levels of the university, a larger sample of classes would provide greater credence to the conclusions drawn.

The need to expand this study to other areas involving perceptions of one person by another is clear. The issue of conditioning of students toward the category scaling has been addressed. Specifically, a study comparing the reactions of students who have had practice and/or warmup on the multiply-divide procedure compared to students who are experiencing the technique for the first time would be helpful, since the written comments from some students indicated that they did not understand the multiply-divide method of responding.

The multiply-divide procedure could also be used to measure changing student attitudes toward teaching. The scale of one million is potentially more sensitive to shifts in students' attitudes, values, and perceptions than the more conventional five point category scale.

Information on the programming of calculations for the multiply-divide procedure and programming to graphically depict the distributions with the CALCOMP PLOTTER are available from the authors.

References

Ekman, G. "Discriminal Sensitivity on the Subjective Continuum." Acta Psychologia, 1956, 12, 233 - 243.

Finnie, B. and Luce, R.D. "Magnitude-Estimation, Pair-Comparison and Successive-Interval Scales of Attitude Items." Department of Psychology University of Pennsylvania Memorandum, 1960, MP-9.

Lindsley, O.R. Handbook of Precise Behavior Facts. Kansas City: Precision Media, 1971.

Pennypacker, H.S., Koenig, C.H., and Lindsley, O.R. Handbook of the Standard Behavior Chart. Kansas City: Precision Media, 1972.

Stevens, S.S. "On the Psycholphysical Law." Psycholgical Review, 1957, 64, 153 - 181.

Stevens, S.S. "A Metric for the Social Consensus." Science, 1966, 151, 530 - 541.