Journal of Criminal Law and Criminology

Volume 66 | Issue 2 Article 8

1975

Scaling Seriousness: An Evaluation of Magnitude and Category Scaling Techniques

George S. Bridges

Nancy S. Lisagor

THE JOURNAL OF CRIMINAL LAW & CRIMINOLOGY
Copyright © 1975 by Northwestern University School of Law

Vol. 66, No. 2
Printed in U.S.A.

SCALING SERIOUSNESS: AN EVALUATION OF MAGNITUDE AND CATEGORY SCALING TECHNIQUES

GEORGE S. BRIDGES* AND NANCY S. LISAGOR**

Unidimensional attitude scaling in social research encompasses a variety of measurement techniques. A relevant issue in the application of any of these, however, is the extent to which different procedures yield similar results. For example, one might expect that since magnitude and category scales represent two distinct and different types of scaling, each would generate different sets of results. In categorical analysis, the subjects' judgements involve placing responses into intervals or categories; magnitude scaling involves judging strength or salience and order on a more expansive and continuous scale. Research in psychophysical scaling suggests, however, that category and magnitude scales are logarithmically related.[1] The purpose of this present report is to empirically display the relationship between these scale types within the context of the measurement of delinquency. It is thought that insight on selecting scale types for delinquency research can be gained by examining whether similar seriousness scores result from these methods.

In 1964 Sellin and Wolfgang developed a seriousness index of delinquent events[2] wherein they asserted that the social harm caused by delinquency should be measured by scaling attitudes. It was argued:

> The criteria for determining degrees of s-riousness must ultimately be determined by someone's or some group's subjective interpretation. If weights were assigned by a few criminologists engaged in the task of constructing a mathematical model, we should regard this as an arbitrary determination. But if judgments were elicited from theoretically meaningful and large social groups, consensus might produce a series of weighted values that meaningful and large social groups, consensus would have validity. . . . Although no external objective criteria, beyond people's judgments, exists for producing a continum of seriousness of delinquent acts, there are objective methods of measurement which have been developed into psychological 'laws' relating two different kinds of psychological scales, these methods can be applied to such nonphysical dimensions as the gradual seriousness of deviant behavior.[3]

To determine the scalability of their stimuli, Sellin and Wolfgang empirically compared category and magnitude scales. The latter technique was then selected for developing the seriousness index. Their choice of method was based on the presumed theoretical strengths of magnitude over category scaling, rather than on any differences that were produced in the scale data. This decision is the focal point of analysis in this paper. First, however, a brief review and discussion of scaling techniques is necessary.

The unidimensional scaling methods that have been developed thus far serve at least two important measurement functions. The most common of these functions involves the construction of empirical indices that describe the strength and direction of individual and group attitudes. By these methods, data can be analyzed with a focus on the attitudinal composition of either an entire sample or any of the

* George S. Bridges is a Research Assistant at the Center for Studies in Criminology and Criminal Law, University of Pennsylvania, Philadelphia, Pennsylvania.

** Nancy S. Lisagor is an Assistant Professor at the Department of Sociology at Stockton State College, Pomona, New Jersey.

The authors wish to express their gratitude to Professors Marvin E. Wolfgang and Phillip C. Sagi for their advice and support. Also, thanks go to Robert M. Figlio for the use of data from his dissertation. This paper grew out of a project undertaken by Ms. Lisagor in a graduate seminar at the University of Pennsylvania.

[1] *See* A. Shinn, *Relations Between Scales,* in MEASUREMENT IN THE SOCIAL SCIENCES (H. Blalock ed. 1974). Shinn gives a comprehensive discussion of this literature and the scaling principles related to this logarithmic transformation of the scale scores.

[2] T. SELLIN & M. WOLFGANG, THE MEASUREMENT OF DELINQUENCY (1964).

[3] *Id.* at 237.

individuals within it. A second function, which is of specific interest to this research involves estimating the magnitude of questionnaire or interview stimuli. The focus here, as developed by Thurstone and later greatly extended by Stevens, is directed towards locating estimates of stimuli on an empirical range of strength or salience. Shinn, in his development of a scale typology, suggested that these stimulus-centered methods could be classified into two types: magnitude and category scales. The former amounts to a continuous ratio scale, a numercial range with an absolute zero point. Subjects utilizing a magnitude scale evaluate the magnitude of stimuli by assigning them scores that represent points on a psychological scale. Category scaling, as primarily developed by Thurstone, involves constructing a rank-ordered continuum of stimuli. In the most common technique of successive categories, subjects are presented with a questionnaire item or stimulus, and then are directed to locate it along a scale of successive categories (usually ranging from low to high) that best describes stimulus magnitude. Statistical estimates of the "true" stimuli magnitudes for both magnitude and category techniques are then derived from sample distributions of such judgments. In effect, scales of the stimuli, as well as estimates of their magnitude, are thus obtained.

It has been argued and shown by Helm,[4] Ekman,[5] Ekman and Kuennapas,[6] and Stevens[7] that the log-linear relationship connecting magnitude and category scales is dependent upon the dispersions of the judgments made by subjects. If judgments on the category scale are homoscedastic (have equal variance) for all stimuli, and the dispersions of judgments on the magnitude scales are heteroscedastic and directly function with scale value, then a logarithmic transformation of the magnitude judgments will lead directly to the category scores. In this vein, Sellin and Wolfgang initially investigated whether this psychophysical property held for their non-physical seriousness stimuli. They employed Helm's principle of equal precision or dispersion by standardizing the variances of their rater's catetory judgments for each stimulus. This effectively transformed the non-linear plot of the category and log-magnitude mean scores into a linear one. In so doing, the logarithmic relation between the scales was confirmed.

This finding has at least two important implications for measuring the seriousness of delinquency. Perhaps most importantly, it suggests that seriousness can be measured psychophysically, as a response to physical stimuli would be. Secondly, the log-linear relationship evidenced in this experiment between scales confirms that the scale scores are a log-linear function of one another. Thus, log-magnitude and equal variance category scores with, of course, some measurement error, are directly related ways of estimating the seriousness of delinquent events. The properties of psychophysical scaling developed by Stevens and others are, therefore applicable to this topic in criminological research.

A second aspect of Sellin and Wolfgang's scale development, however, is their final choice of scale design. Despite the direct logarithmic relation between the methods, it is argued that the magnitude scales have added dimensions of validity that extend beyond this log-linear function. Since the scale values are determined by the subject rather than the experimenter, as is the case in the category design, Sellin and Wolfgang argue that magnitude estimation better taps the "true" psychological effects of stimuli. Secondly, it is implicitly argued that the breadth of range in magnitude scales, being so much greater than that of category scales, enhances accuracy in measurement by providing "intrinsically more information" about judgments. Due to these advantages, Sellin and Wolfgang chose to select magnitude scaling over the category technique for the index construction of seriousness.

The focus of this present research is to demonstrate the similarity between magnitude and category scales of seriousness. The following analysis is directed initially to the log-transform property that ties category and magnitude

[4] Helm, Messick & Tucker, *Psychological Models for Relating Discrimination and Magnitude Estimation Scales,* 68 PSYCHOLOGICAL REVIEW 121 (1961).

[5] Ekman, *Measurement of Moral Judgment: A Comparison of Scaling Methods,* 15 PERCEPTUAL AND MOTOR SKILLS 3 (1962).

[6] Ekman & Kuennapas, *Scales of Aesthetic Value,* 14 PERCEPTUAL AND MOTOR SKILLS 19 (1962).

[7] S. STEVENS, PSYCHOPHYSICS AND SOCIAL SCALING (1972).

scores, and it will compare the distributions of magnitude and category scale data for a range of offense stimuli. It is thought that this approach, differing from the mean score plot that is commonly evidenced in the psychophysical literature, permits a closer examination of the relationship between the scales. Rather than focusing simply on mean stimulus scores, it was intended that the research examine also the shapes of the distributions for each stimulus type. The extent to which these distributions "fit" one another will then be tested statistically between the scaling methods.

A second consideration given in the analysis involves the implications for measuring delinquency that directly arise from the results obtained in the study. If the expected logarithmic relationship between category and magnitude scales is demonstrated, then consideration in measuring delinquency's seriousness should be given to any differences that extend beyond the log transform. Sellin and Wolfgang's concern about intrinsic differences in validity would, therefore, be justified. Otherwise, any empirical differences that arise in the analysis should be examined more closely and Sellin and Wolfgang's considerations re-evaluated.

Method

Data were drawn from an analysis of offense seriousness and respondent background made by Figlio.[8] In one part of Figlio's research, two samples of subjects judged the seriousness of twenty delinquent events; one group employed a magnitude estimation scale (N = 158) and the other a category scale (N = 58). Subjects were given test booklets similar to those used by Sellin and Wolfgang.[9] Each booklet contained a thorough set of instructions as well as a list of the twenty delinquent events to be judged. Data generated from these judgments served as the basis for comparing the scaling techniques.

Samples

The samples were drawn from undergraduate sociology classes at the University of Pennsylvania. As Sellin and Wolfgang noted, "most students enroll for the course in introductory sociology without any special factor of student enrollment that might cause them to cluster in a particular way."[10] Thus, it was assumed that the students represented samples from a larger student population that would be free of any unusual attitudinal or intellectual qualities.[11] It was further assumed that the multitude of factors that contribute to students' decisions about which course offering to take, as well as which class times and the teachers preferred in each course, necessarily added to this representativeness and also introduced independence between the sample groups. At a large university with many course offerings, this assumption seems reasonable.

Since the purpose of our analysis was merely to demonstrate scale differences rather than make generalizations about the perceived seriousness of delinquency, we proceeded with comparisons and tests of the scaling techniques. Any other biases that may have been operant in the data due to sampling and measurement error were beyond the control of the investigators.

Measures

Figlio provided a concise description of the category and magnitude scales that best suited the discussion to follow. He noted:

> In the category scale each subject was asked to circle the number from one to eleven (least to most serious) which best represented how serious he thought that particular offense was. In the magnitude scale, the subject was asked to choose any number which adequately represented the seriousness of that particular offense description. The category scale has the advantages of being easy to visualize and to understand, it is also numerically constraining. The magnitude scale, while having no such constraint, requires greater abstraction in the thought process.[12]

[8] Figlio, *The Seriousness of Offenses: An Evaluation by Offenders and Nonoffenders,* 66 J. CRIM. L. & C. 189 (1975).

[9] T. SELLIN & M. WOLFGANG, *supra* note 2, at 237.

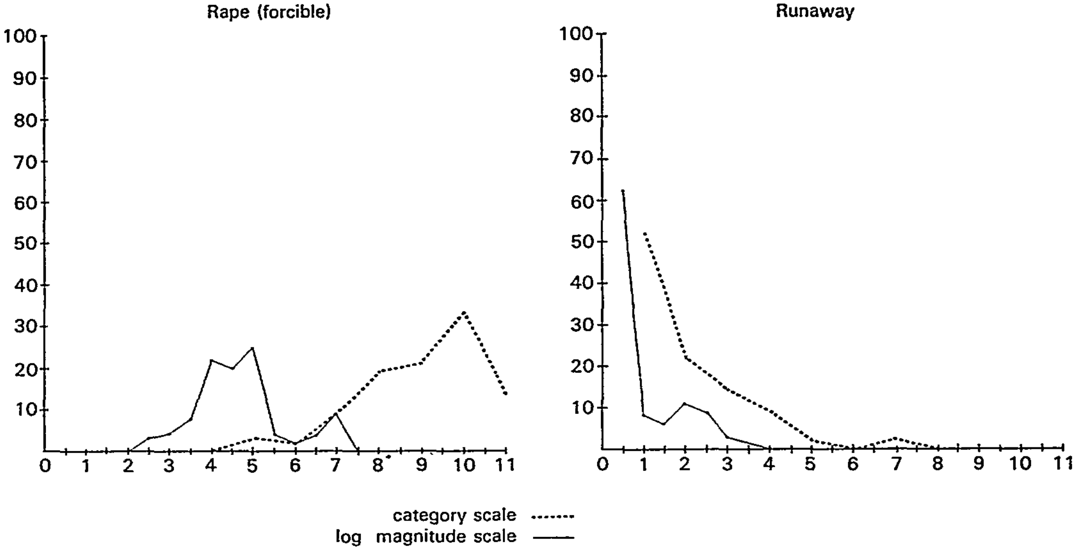
[10] *Id.*

[11] As noted previously, the assumption of homoscedastic or equal dispersion category scores underlies the log-linear transform linking category and magnitude scale scores. An examination of the sample category scores indicated that the dispersions around the twenty offense stimuli were approximately equal. Also, regression analyses of the mean category scores (raw) on the mean log-magnitude scores gave a very tight linear fit. By these findings, it was assumed that the property of homoscedasticity was operant in the category scale.

[12] Figlio, *supra* note 8.

FIGURE 1

RAW CATEGORY AND LOG-MAGNITUDE SCALE



* Frequency distributions in percentages
(abscissa is scale value, ordinate is percentage value)

Particular consideration was also given to biases introduced by stimulus and response effects on the judging process. As Sellin and Wolfgang assert: "A larcency following a disorderly conduct might be judged more serious than a larceny following a murder; a charge of intoxication might be judged more serious if it followed a runaway offense than if it were preceeded by a larceny; and so forth."[13] In order to reduce these potential distortions, the same items were used for each scaling technique and were randomly assigned positions in the test booklet. Thus, the final measurement tool was two identical sets of randomized offense descriptions, one that had category response scales and the other with boxes for magnitude scoring.

Analysis

Our primary concern was to establish the effects of the logarithmic transform property on category and magnitude scale distributions. Four steps were taken to meet this goal: 1) the logarithms of the magnitude scale data were calculated,[14] 2) a transformation was made on the category scale data so that the discriminatory dispersions of each subject were made equal, 3) the log-magnitude scores per subject were standardized so that comparisons between the distributions could be made on a single numeric continuum and, 4) the strength of the relationship between the scale data for each stimulus was examined statistically.

Results

The frequency distributions of the log-magnitude and raw category scores for two of the twenty offense stimuli, forcible rape and running away from home are found in Figure 1.[15] From simple inspections of the graphs, the similarity between the scale curves is evident in their general shapes. For most offenses, the

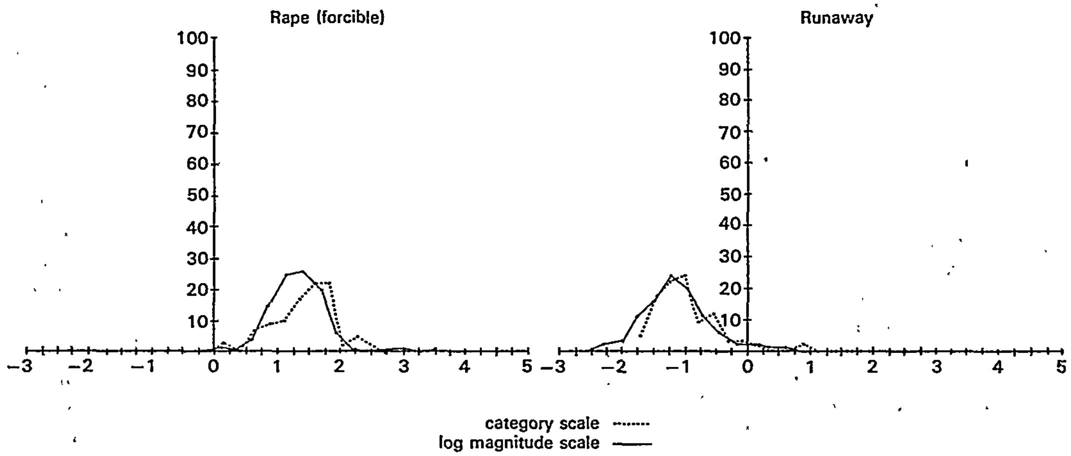
[13] T. Sellin & M. Wolfgang, *supra* note 2, at 253.

[14] Naperian or "natural" logarithms were taken in the transformation of the magnitude scores.

[15] The presentation of the entire set of scale distributions was saved for the sake of brevity in the report. Distributions for forcible rape and running away from home were selected because they are representative of other serious and non-serious offense stimuli. Copies of the entire set of distributions are available from the authors upon request.

FIGURE 2

RAW CATEGORY AND LOG-MAGNITUDE SCALE*



* Frequency distributions in percentages
(abscissa is scale value, ordinate is percentage value)

modal levels of the distributions were found to be approximately equal. It seemed, however, that one distinguishable difference between the distributions was location. The category scale scores were consistently located at points higher on the seriousness scale than the magnitude scores. It was also evident that the category distributions had slightly higher dispersions than the magnitude distributions. These differences indicated that the log-magnitude and category scores, although obviously correlated by their similarity in shape, differed by location and scale factors. It should be noted, however, that this was not inconsistent with Sellin and Wolfgang's findings. The relationship they demonstrated between the scales was log-linear; category scales were related to magnitude scales by multiplicative and additive constants. This suggests that the scale distributions should have differed only as a function of these. constant terms. Thus, a simple transformation of the data to eliminate these constant effects was sought in order to render the distributions comparable and equivalent.

To determine the nature of this transformation, the study was next directed toward the variation in the judgments of individual subjects. It was thought that unless the variation between individual subjects in the way each judge stimuli is taken into account, the variation around each stimulus would differ in a misleading manner. In order to standardize this subject effect, location and scale transformations were made on each subject's evaluations. This amounted to transforming the data such that all had mean scores of zero and equal variances. These standardized judgments, plotted for forcible rape and running away, are found in Figure 2. The marked similarity between the curves then becomes even more apparent; it is evident that, by performing a simple standardization of the log-magnitude and category scale data, the expected logarithmic relationship between the scale types is revealed. The scales produced extremely similar results across all of the stimuli under measure.

In order to describe this relationship statistically, chi-square goodness-of-fit tests were performed on each of the offense stimuli. This statistical method involves measuring the extent to which two frequency distributions differ at points along a common numerical scale.[16] As is evident from Table I, two of the

[16] One important consideration that must be made when using the chi-square goodness-of-fit test relates to the number of intervals into which data are grouped. Differences between distributions may vary directly with this number. Thus, whether an hypothesis is accepted or rejected is often dependent upon the number of degrees of freedom associated with the test. In order to free

TABLE I

TESTS OF SCALE DIFFERENCES FOR TWENTY OFFENSE TYPES

Offense Stimulus	X^2, degrees of freedom		Probability
Larceny, $5	3.67	6df	p > .20
Larceny, $20	5.78	5df	p > .20
Larceny, $50	13.25	7df	p < .10
Larceny, $1000	9.50	6df	p < .20
Larceny, $5000	7.35	6df	p > .20
Burglary, $5	13.95	7df	p < .10
Robbery (no weapon) $5	7.52	7df	p > .20
Robbery (weapon) $5	5.73	6df	p > .20
Assault (death)	14.38	7df	p < .05*
Assault (hospitalization)	11.80	6df	p < .10
Assault (theft & damage)	6.33	6df	p > .20
Assault (minor)	6.70	6df	p > .20
Rape (forcible)	8.85	7df	p > .20
Auto Theft	27.70	7df	p < .05*
Rifle (no permit)	5.52	7df	p > .20
Trespassing	14.04	7df	p < .10
Illegal Liquor	6.28	5df	p > .20
Disorderly Conduct	8.85	5df	p < .20
Runaway	10.10	6df	p < .20
Hookey	10.70	5df	p < .10

* Significant beyond conventional levels of probability.

twenty distributions differed beyond the conventional levels of probability used in hypothesis testing. Since it was decided to reject the null hypothesis of no significant differences between the distributions at the .05 level of significance, one would expect that one of the twenty tests (.05) would be rejected solely due to sampling error. The data, however, indicate that the scales differ for two offense stimuli: assault resulting in the death of the victim and auto theft. In explanation, one might argue that magnitude and category scales differ as a function of stimulus magnitude or offense seriousness. This appears consistent with the scale data because the magnitude distributions had slightly lower dispersions than the category distributions for serious offenses. This "peaking" effect, however, contributed to only two significant differences between the scales in the twenty comparisons. Thus, it is impossible to be confident statistically that these differences represent differential stimulus effects.

Although these differences are unexpected in light of the scaling principles that have been presented thus far, the two scale distributions closely approximate one another across most of the offense stimuli. This similarity in the distributions supplements our understanding of the differences between the techniques. It is shown, through simple transformations of the scale data, that comparable distributions of seriousness scores are produced. It is important to note that these transformations in no way violate the study's earlier assumptions about the discriminatory dispersions around stimuli. This analysis has involved no more than a shift of the scale axes so that their distributions could be more easily examined.

CONCLUSION

This study has illustrated that magnitude and category scaling techniques produce quite similar distributions and estimates of seriousness magnitude. Although the analysis represents empirically no more than a further validation of a fundamental principle in attitude scaling, the implications of its findings for delinquency research are important. Initially, these results indicate that rigor spent on choosing a "best" scale type for measuring seriousness does not significantly affect the final research product. Regardless of the technique employed (category or magnitude), quite similar findings result. It should be noted, however, that there are some intrinsic differences between magnitude and category scales that warrant consideration. These differences beyond the log transform cannot be directly resolved by simple comparisons of data. The properties of each method must be more closely examined. If the properties of one technique are more directly suited than another to a specific measurement problem in delinquency research, then this technique should be employed. Inherent differences in validity as well

the analysis from this source of bias, the scale distributions were divided, wherever possible, into similar numbers of groups. It was thought that this procedure, while introducing an element of consistency in our method, would best reveal any differences existing between the scale distributions.

as in the efficiency of administration and analysis between the scale types should be considered in this choice of method.

A second area to which these findings relate is the research process itself. A common problem in many areas of social research involves employing measurement techniques that are best suited to research needs. As research grows in analytic complexity, it becomes increasingly more difficult to ensure this suitability. By incorporating comparative analyses of different techniques into specific substantive areas, the suitability of these techniques can be better established. The above finding that the different scaling methods produce similar results provides a validation of psychometric research; the log-linear relation between the scales had been demonstrated by simple algebraic transformations of the scale data. More importantly, it assures the criminologist that magnitude and category scales provide similar estimations of delinquency's seriousness. Both methods are suited to their research problems; each, however, has properties that are appealing to specific research needs and must be considered in future analyses.