ELI P. COX III*

A conceptual framework employing the distinction between stimulus-centered and subject-centered scales is presented as a basis for reviewing 80 years of literature on the optimal number of response alternatives for a scalo. Concepts and research from information theory and the absolute judgment paradigm of psychophysics are used. The outhor reviews the major factors influencing the quality of scaled information, points out areas in particular need of additional research, and makes some recommendations for the applied researcher.

The Optimal Number of Response Alternatives for a Scale: A Review

The debate about the optimal number of response alternatives for a scale virtually spans the history of such instruments. Boyce (1915, p. 20) reviewed the number of alternatives employed in scales used to evaluate the efficiency of teachers. Conklin (1923) argued for the use of the "scale of values method" over conventional "questionnaire studies" (i.e., those employing unstructured-response questions), and lamented that studies using scales went back as far as 1909 but that nc attempt had been made to examine the number of response alternatives systematically. He recommended the use of nine positions over 13 because of the fact that some of the options were virtually neglected when the more refined scale was employed. Symonds (1924) introduced the criterion of reliability (inter-rater correlation), precipitating additional debate and research (Pemberton 1933; Champney and Marshall 1939).

More recent studies have broadened the range of criteria to include test reliability and validity (Cronbach 1950; Bendig 1953, 1954a; Komorita 1963; Komorita and Graham 1965; Jacoby and Matell 1971; Matell and Jacoby 1971); response time (Bricker 1955; Behar 1963; Bevan and Avant 1968; Matell and Jacoby 1972); respondent preference (Jones 1968); usage of the "uncertain" category (Rundquist and Sletto 1936; Ghiselli 1939; Guest 1962; Matell and Jacoby 1972); information theoretic measures (Bendig and Hughes 1953; Bendig 1954; Garner 1960; Peterson and Sharma 1977a); recoverability of synthetic data (Green and Rao 1970; Lissitz and Green 1975; Jenkins and Taber 1977); the interpretability of descriptive statistics (Morrison 1972; Martin 1973, 1978; Martin, Fruechter, and Mathis 1974; Peterson and Sharma 1977b; Best et al. 1978); and the statistical efficiency of sample estimates (Lunney 1970; Benson 1971; Connor 1972; Ramsey 1973).

If the number of response alternatives were to be established democratically, seven would probably be selected (Symonds 1924; Morrison 1972; Ramsey 1973); seven is the modal number of response alternatives for the scales reviewed by Peter (1979). However, others (Peabody 1962; Lunney 1970; Jacoby and Matell 1971) suggest that as few as two or three alternatives may be appropriate under some circumstances. In contrast, Guilford (1954) suggests that the optimal number can be as high as 25, and one information theorist (Garner, 1960, p. 352) found that information transmission increased up to 20 response alternatives (the largest number tested) and concluded that ". . . it is clear that information cannot be lost by increasing the number of rating categories."

There is also a substantial discrepancy on this point among the most popular scaling methods. In the equal-appearing intervals scaling method (Thurstone 1928; Thurstone and Chave 1929), judges are allowed 11 categories in scaling attitude statements and respondents are allowed two alternatives for responding to each statement. The summated scale (Likert 1932; Murphy and Likert 1938) provides respondents with

* Eli P. Cox III is Associate Professor, Department of Marketing, Graduate School of Business, University of Texas at Austin.

Journal of Marketing Research
Vol. XVII (November 1980), 407-22

five alternatives for each statement and the semantic differential (Osgood, Suci, and Tannenbaum 1957) provides seven. Justification for the choice appears only for the semantic differential: five alternatives tend to frustrate college students whereas some alternatives tend to be underutilized when as many as nine are provided.

The purpose of this article is to review the research on the optimal number of response alternatives for a scale. The first section outlines a theoretical framework which should aid in reviewing what is approximately 80 years of research. Torgerson's (1958) distinction between stimulus-centered and subject-centered approaches to scaling is used as a part of this framework and a section is devoted to each approach. In the concluding section the author reviews the major factors influencing the quality of scaled information, points out areas in particular need of additional research, and makes some recommendations for the applied researcher.

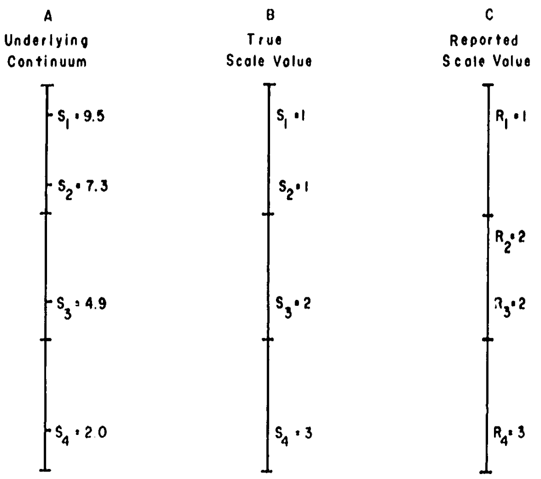
A CONCEPTUAL FRAMEWORK

At a general level, a scale with the optimal number of response alternatives is refined enough to be capable of transmitting most of the information available from respondents without being so refined that it simply encourages response error. At that optimal number, the ratio of meaningful, or systematic, variation to total variation is maximized. At an operational level, the optimal number depends on the purpose of the scale and thus the nature of its systematic variation.

Torgerson (1958, p. 46) presents a tripartite classification of scaling techniques based on the nature of their systematic variation. In the *stimulus-centered or judgment approaches*, the systematic variation in responses is attributed to differences within a set of stimuli with regard to some shared attribute. The data collected from such techniques can be viewed in the format of one-way analysis of variance with the columns corresponding to the stimuli and the rows corresponding to respondents or the replications for a single respondent. *Subject-centered approaches* attribute the variation in responses to the stimuli as actually being due to differences among respondents. These data can also be represented in an analysis of variance format with the columns corresponding to the respondents and the rows corresponding to the stimuli which can be viewed as replications. The *response approaches* accommodate differences in respondents and stimuli simultaneously and their data can be represented in a two-way analysis of variance format with a single observation in each cell. In all three cases, a good scale is one which maximizes the ratio of systematic to random variation.

Although researchers are often interested in the differences among both stimuli and respondents, the response approach is used rather infrequently in marketing. All of the research evaluating the optimal number of response alternatives has involved either stimulus-centered or subject-centered scales.

Figure 1
MAPPING OF STIMULI INTO THREE RESPONSE CATEGORIES



Stimulus-Centered Scales

In the stimulus-centered approach, respondents are generally presented with a set of scale items where each item is used to measure the extent to which one of the stimuli has one of the attributes of concern. For example, a set of 25 semantic differential scales might be used to evaluate five objects on the basis of five attributes.

It seems that the optimal number of response alternatives for a stimulus-centered scale is determined for a particular set of circumstances by (1) the information transmission capacity of the scale, (2) the ability of respondents to assign response alternatives to stimuli, (3) the amount of information available from the stimuli for transmission, and (4) the information needs of the researcher.

Certain aspects of this process are suggested by considering Figure 1 which illustrates the mapping of stimuli into response categories. Column A represents the underlying continuum along which the stimuli of interest are positioned. The actual values assigned to stimuli can be considered direct measures of some objective attribute such as weight or the scale values of some subjective attribute such as "attractiveness" as determined by a scaling technique such as paired comparisons. The variation among these scores represents the maximum amount of information available for transmission.

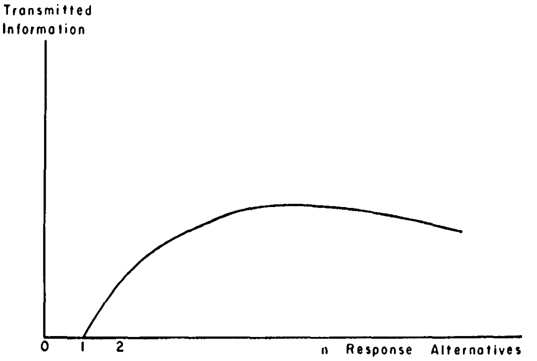
Column B represents the stimulus measures obtained after the continuous data have been accurately converted into class-interval data, based in this case on three intervals. This conversion process causes a loss of information. For example, no difference is indicated between S_1 and S_2 in column B, whereas S_1 has significantly more of something if the underlying continuum is at least interval-scaled. It is argued that the extent of such information loss is determined by the coarseness of the scale employed. Alternatively, the information transmission capacity inherent in a scale is determined by the precision allowed by its response alternatives.

Column C represents the scale values assigned to stimuli by respondents. Whereas the translation of continuous measures into class-interval ones introduces imprecision, the translation of class-interval measures into scale responses may introduce inaccuracy. For example, R_2 should have been assigned the value of one rather than two.

It seems that as the number of response alternatives is increased beyond some minimum, the demands placed upon a respondent become sufficiently burdensome that an increasing number of discrepancies occur between the true scale value and the value reported by the respondent. Thus, though the information transmission capacity of a scale is improved by increasing the number of response alternatives, response error seems to increase concomitantly. Accordingly, one might hypothesize that the relationship between the amount of information actually transmitted by a scale and the number of response alternatives it provides is similar to that shown in Figure 2.

Probably several factors mediate this relationship, but the most important seems to be the amount of information available among the stimuli in the first place. Clearly, if the stimuli are identical, or if the differences among them are trivial, only a single response category is necessary. If the perceived or actual differences among the stimuli are large, a correspondingly refined scale is required to represent these differences. However, a scale refined beyond the level necessitated by the stimuli simply encourages reponse error. Thus, the point at which the transmitted information begins to decay seems to be determined by the homogeneity of the stimuli.

Figure 2
RELATIONSHIP BETWEEN THE AMOUNT OF INFORMATION TRANSMITTED BY A SCALE AND THE NUMBER OF ITS RESPONSE ALTERNATIVES



The final factor to be considered in this conceptual framework is the amount of information required by the researcher. It might be argued that the optimal number of response alternatives is found at the point where the amount of transmitted information is maximized. However, the researcher may not find this amount of information cost-beneficial or even necessary. Thus, a researcher may simply be interested in knowing whether respondents will buy a product or not, even though respondents may be capable of representing the strength of feeling underlying that action or inaction with great exactitude. Very often, the hypotheses we examine in marketing require only the crudest of measures; the information-processing capacity of the researcher rather than that of the respondents may dictate the optimal number of response alternatives in many cases.

Subject-Centered Scales

Generally, subject-centered scales are composites of a number of items which have been selected to position respondents along a continuum representing a single attribute. The information transmission capacity of such composite scales is a function of the number of items as well as the number of response alternatives. For example, a composite scale consisting of five Likert scale items with five response alternatives would allow for 3,125 (I^R) unique patterns of response. Only 21 ($IR - (I - 1)$) total scores are possible from such a scale because many of the response patterns are scored similarly (e.g., 381 patterns of response will produce a score of 15).

Test reliability and validity are the primary criteria used to evaluate subject-centered, composite scales. In terms of analysis of variance, the items are replications which combine to provide a reliable scale if the ratio of between-respondent variation to total variation is high. The scale is valid if the respondents' mean values are at least monotonically related to their true positions on the attribute continuum.

It seems that the reliability of a scale would be influenced by several factors. First, the relationship between the reliability of individual items in a composite scale and the number of response alternatives would be similar to that of transmitted information for stimulus-centered scales, shown in Figure 2, and for the same reasons. Second, the addition of meaningful scale items should increase the information transmitted by the scale and thus its reliability. Third, the meaningfulness of a set of items can be viewed as the extent to which they measure the same attribute. Thus, the reliability of a scale should be directly related to the covariation of its items.

OPTIMAL NUMBER OF RESPONSE ALTERNATIVES FOR STIMULUS-CENTERED SCALES

Three traditions of research are discussed in this section. In the first tradition, information theory has been utilized to evaluate the amount of information that is transmitted by a scale as the number of response alternatives is varied. In the second, the absolute judgment paradigm of psychophysics has been employed to evaluate the information processing capacity of human judges and the results have been considered by some researchers to be relevant to scale design. Third, metric techniques of analysis have been employed in addressing this question. Although this is the oldest of the three traditions of research, it is discussed last in order to benefit from the concepts introduced in information theory.

Information Theory

Obviously, our ability to test the hypothesized relationship between the amount of information transmitted by a scale and the number of its response alternatives depends on our ability to measure information. Shannon and Weaver (1949) are generally credited with popularizing the view that information is quantifiable. Miller and Frick (1949), Garner and Hake (1951), and McGill (1954) demonstrated the relevance of this information theory to psychology and presented methodologies for approaching psychological problems. An excellent overview is provided by Attneave (1959).

A bit (binary unit), the unit in which information is expressed, is equivalent to the amount of information required to distinguish between two equally probable events. According to this line of reasoning, if a respondent were equally likely to choose any one of the responses on an eight-point rating scale or a multiple choice question with eight options, three bits of information would be required to identify the actual response chosen ($\log_2 8 = 3$).

Five nominal-scaled information measures have been used in evaluating the optimal number of response alternatives for a stimulus-centered scale. "Stimulus information," $H(x)$, measures the amount of dispersion among the stimuli being scaled and represents the maximum amount of information available from the stimuli to be transmitted.

H_{max} is a measure of the maximum amount of dispersion, hence information, that can be transmitted by a scale and increases monotonically at a decreasing rate with the number of response alternatives. Coombs (1964) takes an information theoretic view of data collection techniques and refers to this measure as the "channel capacity" of a technique.

"Response information," $H(y)$, is a measure of dispersion in the responses obtained by means of a scale and as the number of response alternatives increases it should increase but at a slower rate than H_{max}.

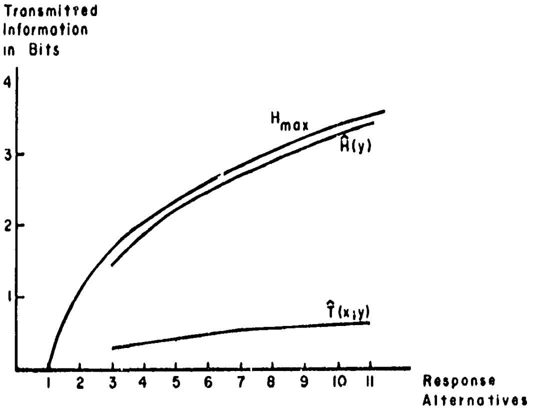
$H(y)$ is composed of the amount of "transmitted information," $T(x;y)$, and an error term, $H_x(y)$. $T(x;y)$ is a measure of the association between the nominal-scaled measure of dispersion for the stimuli, $H(x)$, and the nominal-scaled measure of dispersion for the responses, $H(y)$, and is similar to the contingency coefficient. According to the conceptual framework presented before, one would expect that $T(x;y)$ would always be less than $H(y)$ and would increase with the number of response alternatives until the optimal is reached, then level off or decline.

Although Garner and Hake (1951) suggest that an information theoretic approach could be used in evaluating scales, the first researchers to do so appear to have been Bendig and Hughes (1953). They asked 225 college students to rate 12 countries according to how much they knew about them. The respondents were assigned to one of 15 experimental conditions, each representing a different combination of number of response categories (3, 5, 7, 9, 11) and number of verbal anchors (end points, midpoint, all three). Figure 3 shows the results of this analysis. The values for the experimental conditions representing the number of verbal anchors are averaged to provide the values for each number of response categories. $\hat{H}(y)$ increases directly with increases in the number of alternatives. However, $\hat{T}(x;y)$ rises only from .36 to .59 and no decline is evident.

In a second study, Bendig (1954) obtained preference ratings for 20 foods from 236 respondents who were assigned to experimental conditions representing the number of response alternatives (2, 3, 5, 7, 9). $\hat{T}(x;y)$ increased from .06 to .34 and again no decline was apparent.

Figure 3

RELATIONSHIP BETWEEN THE NUMBER OF RESPONSE ALTERNATIVES AND TRANSMITTED INFORMATION FOUND BY BENDIG AND HUGHES (1953)



Garner (1960) asked 50 respondents to evaluate 20 different handwriting samples on a four-point scale. The process was repeated for scales with 6, 8, 10, 12, 16, and 20 response categories for a total of 7,000 ratings. The presentation of both the samples and the scales was varied to eliminate order effects. $\hat{T}(x;y)$ increased approximately linearly from four to 20 categories and the rate of change was slight. Garner employed the corrections recommended by Miller (1955) to adjust for the fact that $\hat{H}(y)$ tends to be biased downward and $\hat{T}(x;y)$ tends to be biased upward. Although these corrections were not applied in the previous studies, the results are consistent with Garner's.

Behar (1963) conducted an experiment in which groups of six subjects evaluated the size of 10 squares using two response categories and 10 categories. Response errc', $\hat{H}_x(y)$, was greater for 10 alternatives than for two, although $\hat{T}(x;y)$ was not reported. It was also found that response equivocation was greater for the largest and smallest squares than for the medium-sized ones. Bevan and Avant (1968) obtained similar results when using 2, 4, 8, 16, 32, and 64 response alternatives to evaluate the size of 10 squares. In addition, they found that $\hat{T}(x;y)$ increased rapidly to eight response alternatives and virtually none thereafter.

Little evidence is found in these studies to support the hypothetical relationship between $\hat{T}(x;y)$ and the number of response alternatives. The inconsistency is both in the modest increases in $\hat{T}(x;y)$ found by everyone but Bevan and Avant and in the fact that it did not decline even for the largest number of response alternatives. Interpreting these results at face value might lead to the conclusions that (1) the optimal number of response alternatives was larger than the number of alternatives explored and (2) increasing the number of alternatives tended to produce only the most modest increases in transmitted information.

However, the optimal number of alternatives is not likely to have been revealed even if the ranges considered in all but the last study had been increased significantly. The reason is that beyond a certain limit an increase in the number of response alternatives becomes meaningless in information theory. The maximum value of $\hat{T}(x;y)$ occurs when a respondent places each stimulus in a separate response category. Under such circumstances, the responses would all be in the diagonal cells of the data matrix. No benefits can be obtained by offering additional response alternatives as was done by Bevan and Avant. In other words, a respondent can transmit no more information than is found in the stimuli in the first place. Consequently, according to information theory, the optimal number of response categories will never exceed the number of stimuli being scaled.

Clearly, this constraint is imposed by the nature of the information measures. These measures suggest that no more than one response category would be required if a single stimulus were being scaled. Actually, all of the information measures would be equal to zero under these circumstances. And, if two or more response categories were provided to "scale" a single object, $\hat{H}(x)$ would remain equal to zero, as would $\hat{T}(x;y)$, and any increase in $\hat{H}(y)$ that would occur would be the result of increases in the error term, $\hat{H}_x(y)$. This special character of information measures seems inappropriate for the present purposes for two reasons. First, though it is unclear that scales are interval in character, they certainly provide more than either nominal or ordinal data. They at least meet the requirements of Torgerson's (1958, p. 16) "ordinal scale with a natural zero" and thus would make the use of several response alternatives meaningful even if only one object is being scaled. For example, a scale evaluating whether respondents find a single soft drink to be sweet or sour should have at least two scale positions.

Second, the fact that the information measures are stimulus-centered limits their application. Information measures based on the individual respondent as the unit of analysis have been developed (see Attneave 1959) but they have not been employed in the present connection.

In conclusion, the nonmetric, multidimensional measures of information enable researchers to compare results across a variety of situations, but this universality is acquired at the cost of imprecision. Certainly, improvement can be made in the application of information theory to the determination of the optimal number of response alternatives. It is not at all clear, however, that the sacrifice of precision is justified as the data are always at least ordinal.

The Absolute Judgment Paradigm

Miller's (1956) excellent review of absolute judgment experiments has been cited in the marketing literature to support the argument that limited benefits are derived from allowing respondents to make fine discriminations when using rating scales (Green and Rao 1970; Morrison 1972; Hulbert 1975). In this section human information processing capacity is discussed in the context of the absolute judgment paradigm and the relevance of this research to the question of rating scales is assessed.

Miller's propositions about human information processing capacity are based on the synthesis of several experiments on the absolute judgments of unidimensional stimuli. The subjects in these experiments are provided a set of stimuli and an equal number of responses (usually numerals) with which to identify the stimuli. A practice period allows the subjects to learn the correct stimulus-response pairings as best they can. The actual experiment involves presenting the stimuli at random and asking the subjects to identify them. The independent variable in these studies involves the nature and number of stimuli, and the dependent variable is the number of correct identifications which when expressed in information theoretic terms is in bits.

Typical of these studies is one conducted by Pollack (1952) in which subjects were asked to identify tones of varying frequency by assigning numerals to them. When only two or three tones were presented in an experimental condition, subjects made no mistakes in identifying them. As the number of different tones increased to four, mistakes of identification began to appear and they became increasingly frequent as the number of tones increased to 14. The outcome can be expressed more precisely in terms of information theory: as the number of tones presented increased to 14, $\hat{H}(x) = 3.81$, transmitted information, $\hat{T}(x;y)$, leveled off at about 2.5 bits.

Miller's synthesis is based on Pollack's study as well as others involving (1) loudness, (2) saltiness of a solution, (3) position of a pointer on a linear interval, (4) size of squares, (5) intensity, duration, or position of vibrations on the skin, and (6) curvature, length, or direction of lines. Despite the variety of these unidimensional experimental stimuli, Miller notes a remarkable similarity of findings. The mean across the experiments for the maximum number of stimulus categories that could be utilized successfully by subjects was approximately 6.5 (2.6 bits), with one standard deviation including from 4 to 10 categories (.6 bits). Accordingly, Miller concludes that (1) the notion of channel capacity appears to be valid for describing human observers and (2) the variation in the level of channel capacity appears to be remarkably consistent across the situations examined.

In examining these experimental findings and others presented in his review, Miller underscores several limitations to his conclusions. The most important of these relate to the facts that (1) absolute judgments were employed, (2) unidimensional stimulus objects were investigated, and (3) perception rather than memory was being assessed.

First, the fact that absolute judgments were investigated meant that subjects were presented with a *single* stimulus object at a time and asked to judge it. If subjects had been presented with pairs of stimulus objects and asked to discriminate between them or if they had been asked to rank-order a large set of objects along the appropriate dimension, they would have been able to make finer discriminations more accurately. Although Pollack found that subjects are very limited in their ability to make accurate absolute judgments of tones, he cites Stevens and Davis (1938, p. 152–4) who estimated that young subjects can distinguish among approximately 350,000 tones when they are presented in the form of pairwise comparative judgments.

Second, the studies involved unidimensional rather than multidimensional stimuli. Miller reviews several experiments investigating multidimensional stimuli and concludes that the channel capacity of subjects increases with the number of dimensions but at a decreasing rate. As a consequence, the magic number seven which has been so widely cited increases to approximately 150 categories in a laboratory experiment involving six-dimensional tones. Further, "everyday experience teaches us that we can identify accurately any one of several hundred faces, any one of several thousand words, any one of several thousand objects" (Miller 1956, p. 87).

Miller also suggests that it is probably necessary to distinguish between the "span of absolute judgment" and the "span of perceptual dimensionality," which is concerned with the limit which may exist on the number of common dimensions along which we can evaluate a set of objects. Though the magic number for the former is seven, Miller indicates that it is probably more like 10 for the latter, although the question has not been adequately examined. Much research has been done on this topic since publication of Miller's article and addresses the question of whether more information is necessarily better (e.g., Bartlett and Green 1966).

Third, the studies cited involved the examination of perceptual discrimination in a rather restrictive setting and the accuracy of human judgment may be expanded when mnemonic processes are employed. Coupling human information processing with memory greatly increases channel capacities.

Miller suggests that humans organize, or recode, many small pieces of information into a smaller number of summary "chunks." To make this distinction, suppose that an individual is shown the following series of letters for 10 seconds and then is asked to repeat them: a penny saved is a penny earned. A child who knows only the alphabet would have to retain 25 pieces of information, whereas a child who can read would need retain seven chunks of information and someone familiar with the proverb would need retain only one, still larger, chunk of information. Thus, learning and experience increase our channel capacities.

On the surface, the demands placed on an individual when responding to stimulus-centered scales and when participating in an absolute judgment experiment are very similar in that both activities require stimuli to be represented using a set of response alternatives. In fact, Coombs (1964, p. 221) suggests that there is no distinction between these methods on the level of data theory.

However, the significance of the previous distinctions as caveats when one attempts to extend Miller's findings to scale design is rather obvious. The stimuli generally evaluated via stimulus-centered scales in marketing studies are much more complex than those represented in absolute judgment experiments and the responses are often based on years of experience and learning. Moreover, the ordered responses on a scale, particularly when labeled, have intrinsic meaning not found in numerals designed to differentiate tones or vibrations. In short, it seems possible to greatly underestimate human information processing capacity and thus the ability of respondents to provide fine, yet meaningful, discriminations among stimuli via scales with numerous response alternatives.

One final point must be made about the correspondence between an information theoretic approach to absolute judgments and to determining the optimal number of response alternatives for a scale. Garner (1960; also see Garner and Hake 1951) suggests that the objectives of the two analyses are very different. In absolute judgments, the number of stimuli and the number of responses are varied identically and the objective is to determine the minimum number of symbols (responses) required to transmit information efficiently. The researcher is free to select both the nature and number of stimuli so as to maximize discrimination. Saffir (1937), Attneave (1949), and Garner and Hake (1951) developed methods for so selecting stimuli. In contrast, the researchers employing stimulus-centered scales have a fixed number of stimuli and are free to vary the number of response alternatives on the scale so as to obtain information about each stimulus, even if it is obtained in what appears to be an inefficient manner. This distinction may suggest why the number of useful response alternatives is greater than the measure of transmitted information suggested in the studies discussed previously.

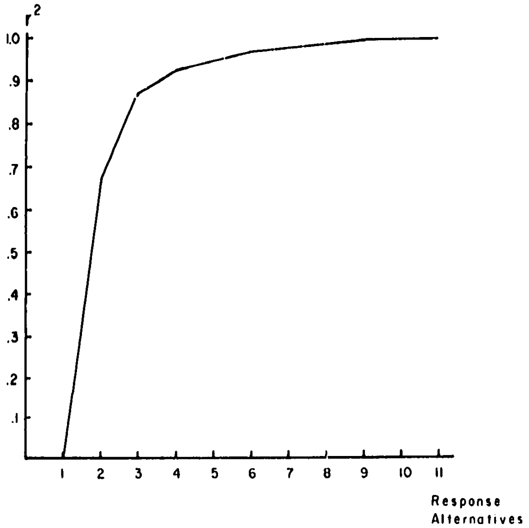
Metric Approaches

The vast majority of studies on the optimal number of response alternatives have employed metric dependent variables. Generally, these studies have used correlational measures reflecting either the recoverability of data or some measure of reliability. Before such studies involving stimulus-centered scales are reviewed, it is useful to consider the channel capacity of a scale in metric terms.

Morrison (1972) examined the extent to which the coefficient of determination is depressed when discrete (class-interval) dependent variables are employed in regression analysis. This work is relevant to the present circumstances because $r^2_{ab} = r^2_{ba}$, as $r_{ab} = r_{ba}$. Figure 4 shows the relationship between the number of response alternatives and the channel capacity of a scale when it is defined as the maximum percentage of variation in the underlying continuum that can be explained. For example, 87.5% of the information available can be transmitted with a three-point scale, whereas 99% can be captured with an 11-point scale.

Both r^2 and $H_{\max}$ undergo monotonic increases as the number of response alternatives increases. r^2 undergoes diminishing returns more rapidly and has an asymptote. Accordingly, increasing the number of response alternatives from four to eight increases channel capacity by approximately 4.5% in terms of the metric measure but by 50% according to the information theoretic measure. Despite the dissimilarity of these measures, both indicate that the information transmission capacity of a scale increases as the number of response alternatives increases and that this capacity is definitely limited with coarse scales.

Figure 4
GAIN IN r^2 OBTAINED BY USING A MORE REFINED SCALE

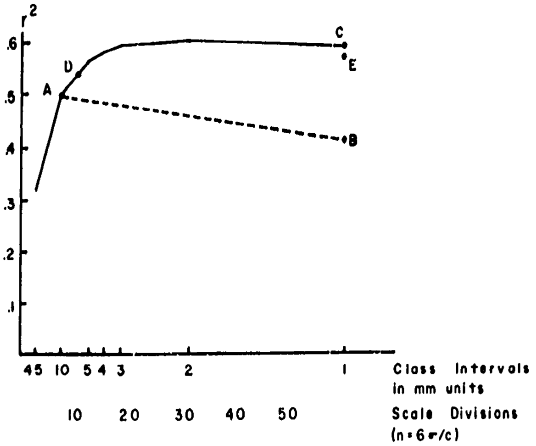


Incidentally, the discovery that descriptive statistics are biased by using class-interval rather than continuous data is not a recent one. Sheppard (1898) presented a correction for variance which tends to be overstated. This would explain why statistical efficiency is reduced by using coarse scales (Benson 1971; Connor 1972; Ramsay 1973).

Kelley (1923) used Sheppard's correction in his own for the product-moment correlation which tends to be understated when class-interval data are employed. This phenomenon has been explored by others more recently (Carroll 1961; Nishisato and Torii 1971; Martin 1973, 1978; Peterson and Sharma 1977b). Also, Martin, Fruechter, and Mathis (1974) found that the conversion of continuous to class-interval data with broad intervals decreases the size of eigenvalues, communalities, and factor loadings in factor analysis.

Symonds (1924, p. 457) states that in evaluating the number of response alternatives, "one has to choose how much of a loss in reliability one will permit in order to make easier the rating." He argues that a person should be more willing to tolerate a loss of reliability when it is low to begin with than when it is high because the statistical reliability of such estimates is also low. Accordingly coarser scales are

Figure 5
RELATIONSHIP BETWEEN INTERITEM r^2 AND THE REFINEMENT OF A GRAPHIC RATING SCALE FOUND BY CHAMPNEY AND MARSHALL (1939)



acceptable when low levels of reliability are found. Using this line of reasoning and Kelley's correction, he identifies the theoretically "ideal" number of response alternatives for various reliability levels and concludes that because the typical reliability measure (inter-rater correlation) is .60, the optimal number of response alternatives is seven.

Champney and Marshall (1939) investigated the optimal degree of refinement that should be used in coding responses obtained with graphic scales. Responses on the nine-centimeter scales were coded by measuring their position to the nearest millimeter such that scores ranged from 10 to 99. One scale was presented in two alternative forms to 80 respondents and r^2 was calculated between the pairs of ratings. The responses on the two scales were recoded to provide four additional sets of class-interval data where the interval ranged from 2 to 45 millimeters in length and r^2 was calculated for each set of data.

The results are represented by the solid line in Figure 5. The x-axis is in terms of the width of the class-interval in millimeters (c) and in terms of the corresponding number of scale divisions ($n = 6\sigma/c$). The apparent lack of correspondence between the two measures is due to the fact that one extreme of the graphic scale was unused and the class-interval scales were assigned to the used portion of the scale. It should also be pointed out that the r^2 values shown are actually mean estimates. Recognizing that the arbitrary location of the class boundaries introduced error in their calculations, Champney and Marshall used several alternative assignments, calculated r^2 for each, and then averaged these values. Consistent with the hypothesis offered previously, average r^2 increases as the width of the interval decreases from 45 to 2 millimeters and then decreases slightly for the 1 millimeter interval.

The authors took the coded responses for the nine-point scale (the centimeter intervals) and added a second digit to them at random to provide new millimeter measures. The r^2 between these new measures is represented as point B in Figure 5. Thus, the decline from point A to point B represents what might occur if nine divisions is optimal and additional refinement goes beyond the respondents' ability to make meaningful discriminations. The fact that point C is above point B suggests that such was not the case in this instance.

As a final step, the authors compared the predictions made by Symonds with their own data. Point D in the figure represents the one place on the curve which corresponds to one of Symond's optimal combinations of reliability (.73) and scale divisions (nine). Point E represents the gain in reliability that would have been achieved if the scale had been refined. Thus, Symond's conceptualization failed to predict the additional gains in reliability that were achieved as well as the fact that reliability would decline after 30 scale divisions.

The data involving respondents' knowledge of foreign nations that were evaluated by Bendig and Hughes (1953) using transmitted information as a criterion were also analyzed in terms of rater reliability by Bendig (1953). A set of scales has rater reliability to the extent that the ratings of stimuli are similar from one respondent to another (see Guilford 1954, p. 395). Rater reliability measures are appropriate for stimulus-centered scales rather than the more conventional reliability measures used to evaluate tests or composite scales which are subject-centered.

Bendig utilized a measure of rater reliability developed by Ebel (1951). This measure can be calculated either for a group of respondents or for the individual respondent. The difference between these measures is an indication of the benefits derived from the use of multiple respondents which are equivalent to those derived from using multiple items in a subject-centered composite scale.

Bendig utilized Ebel's group measure for three groups of five respondents within the experimental conditions representing scales with 3, 5, 7, 9, and 11 response alternatives. The group means (.68, .68, .67, .69, and .65, respectively) are not consistent with the conceptual framework and are not statistically significant. On the basis of this and other analysis, Bendig suggests that the only apparent relationship was that reliability tended to remain unaffected between three and nine categories and declined for 11 categories.

Bendig (1954b) also reanalyzed the data for his food study (Bendig 1953) which was discussed previously. Ebel's measure of reliability for the individual respondent was used as the criterion. This reliability score rose from .07 for two categories to .33 for three, declined to .24 for seven, and remained there for nine categories.

A fourth example of a metric analysis appears in the study of handwriting samples by Garner (1960) which was also discussed in the Information Theory section. The data consisted of seven matrices, each corresponding to one of the experimental treatments (scales with 4, 6, 8, 10, 16, or 20 categories). Each of these matrices was transformed into paired comparisons data by a technique recommended by Guilford (1938). The conventional paired comparisons technique was used to transform these data into a scale corresponding to each experimental condition. Finally, the standard deviation of the stimuli's scale values within each experimental condition was calculated as a measure of the meaningful discrimination obtained. The results of this analysis were similar to those of the analysis of transmitted information: discriminability increased relatively modestly and apparently rectilinearly over the range investigated.

Green and Rao (1970) used synthetic data to evaluate the recoverability of configurations via nonmetric multidimensional scaling and factor analysis when the data were obtained through scales. The independent variables in the 4 × 4 design were the number of response vectors or scales (4, 8, 16, and 32) and the number of response categories (4, 8, 16, and 32). The dependent variable was the correlation between the actual and recovered interpoint distances for a configuration of 18 points.

The correlation increased with both the number of scales and the number of response categories and significant diminishing returns occurred with more than eight scales and more than six categories. In addition, an interaction was found between the variables such that differences in the recoverability of data due to the number of scales were less for many response alternatives than for few. The authors conclude that despite the tentative nature of the results, they should be "heartening" for seven-point scale advocates.

The authors describe some of the limitations of their research, but two points merit notice here. The fact that the data from composite, stimulus-centered scales were analyzed by means of nonmetric multidimensional scaling is of special significance. Metric configurations are obtained from nonmetric distance statements because of the redundancy in those statements which increases with the ratio of the degrees of freedom of the statements to those of the stimulus coordinates (Coombs 1964, p. 39; Young 1970). In this case there were 153 distance statements and 36 coordinates. Possibly the redundancy achieved by increasing the number of points being scaled is akin to that achieved by increasing the number of items in a subject-centered composite scale and thus mediates the relationship under investigation. Also, it should be added that the simulation did not systematically vary the covariation of the scales or response error. The significance of these points is made clear in the next section.

In summary, there is a paucity of research utilizing metric techniques for examining stimulus-centered scales. This lack is particularly surprising when one considers the vast literature of psychophysics. Of the research that has been conducted, only that by Champney and Marshall appears to be entirely consistent with the conceptual framework. Whether the general lack of support is an indictment of the framework or merely indicates the limited scope of the studies in examining a complex phenomenon cannot be determined until much additional research has been conducted.

Summary

All three traditions of research have shed some light on the question of the optimal number of response alternatives for a stimulus-centered scale. The major contribution of information theory has been conceptual. The notions that information can be measured and that the amount of information that can be transmitted by a scale is ultimately determined by its channel capacity are critical to our understanding of this research problem. It is less clear that the information theoretic measures offer any improvements over the more conventional metric measures. Significant insight into the problem can be gleaned from the vast psychological literature. However, extreme caution must be used in applying a generalization made in a limited context, such as the absolute judgment paradigm, to the design of scales. The fact that Pollack's estimates of human information processing capacity for tones are dramatically different from those of Stevens should underscore the fact that Miller's conclusions about that capacity hold true under a very restrictive set of circumstances. Studies employing metric measures have been too limited to have provided much evidence on the optimal number of response alternatives for stimulus-centered scales. However, this tradition probably offers the greatest potential, particularly if future studies have the rigor that has characterized psychophysical experimentation.

OPTIMAL NUMBER OF RESPONSE ALTERNATIVES FOR SUBJECT-CENTERED SCALES

It is ironic that we know very little about how to determine the optimal number of response alternatives for a stimulus-centered scale consisting of a single item when we seem to know much more about subject-centered scales which are composites of several items. The reasons for this disparity are threefold. First, more research has investigated the characteristics of the latter scales. Second, researchers have been able to take advantage of the most basic principle of psychometrics: with the proper assumptions, a lot of poor information can be transformed into a little good information. The third reason, a corollary to the second, is that the literature on composite subject-centered scales indicates several other factors which enhance the reliability of such scales even when we have not been able to determine the optimal number of response alternatives for the items treated singly.

Before the empirical research on composite subject-centered scales is reviewed, it is useful to consider the criterion of reliability and the concept of channel capacity in this context. The reliability of subject-centered scales depends on the extent to which variation in the stimulus-response or item-response matrix can be attributed to differences among respondents. The test-retest, internal consistency, and alternative forms measures of test reliability are all essentially correlational. As such, these measures tend to be depressed to the extent that the channel capacity of the items from which they are calculated is limited.

To illustrate this point, consider the computational formula for Cronbach's (1951) alpha coefficient presented by Peter (1979).

$$\alpha = \frac{k}{k-1}\left(1 - \frac{\sum_{i=1}^{k} \sigma_i^2}{\sum_{i=1}^{k} \sigma_i^2 + 2\sum_{i}^{k} {}_{>} \sum_{j}^{k} \sigma_{ij}}\right)$$

Reducing the number of response alternatives would tend to increase the item variances but would have little effect on the covariances, as suggested by Kelley's correction for the correlation coefficient. Accordingly, alpha tends to be depressed when only a few response alternatives are provided, *ceteris paribus*. Fortunately for the researcher, all things need not remain the same.

Guilford (1954, p. 393) presents the formula for the reliability of a composite score and states that it is a function of the reliabilities of its component scores, their dispersions, their intercorrelations, and the weights assigned to them in the composite. He also points out that the reliability of the composite will be greater than the weighted sum of the component reliabilities to the extent that there are significant component intercorrelations.

The reliability of a composite score is also raised by increasing the number of components or items. Peter (1979, p. 9ff) discusses this feature in the context of the coefficient alpha. Ghiselli (1964, p. 262) uses a modification of the Spearman-Brown formula to indicate the extent to which a test must be lengthened with comparable items to raise the reliability coefficient to some desired level.

Recently, two Monté Carlo studies have been conducted to investigate the relationship between the number of response alternatives, as well as several other characteristics of composite scales, and reliability. Jenkins and Taber (1977) expanded the method of Lissitz and Green (1975) to a 7 × 7 × 4 × 3 design. The independent variables were the number of response alternatives on each item (2, 3, 5, 7, 9, 10, and 14), the number of these items used in the composite (2, 3, 5, 7, 9, 10, and 14), the covariance among the items (.2, .5, and .8), and judgment precision (.50, .70, .85, and 1.00). The complement of judgment precision is the percentage of response variance accounted for by response error, which was generated as an error term in the simulation. Three dependent variables were employed: the squared correlation between observed and true scores, the alpha coefficient, and test-retest correlation (between pairs of samples). A total of 50 samples of 100 scores was generated.

All four main effects made significant contributions to explained variation and the interactions accounted for very little. Beyond this point, the results were divergent for the three dependent variables. For example, item covariance accounted for 50% of the variation in alpha, 13% in test-retest correlation, and only 1% in the squared correlation between observed and true scores. The number of response alternatives ranked fourth in importance of the four independent variables in terms of the reliability measures, explaining 2% of the variation in alpha and 8% in the test-retest correlations. It was ranked third of four for the remaining independent variable, accounting for 14% of its variance.

The mean squared correlations between observed and true composite scores for various numbers of items and response alternatives are listed in Table 1. These results suggest that almost the same degree of accuracy can be achieved by having nine items with two response alternatives as with two items with

Table 1

MEAN SQUARED CORRELATIONS BETWEEN OBSERVED AND TRUE COMPOSITES BY THE NUMBERS OF ITEMS AND RESPONSE ALTERNATIVES FOUND BY JENKINS AND TABER (1977)

	Categories						
Items	*2*	*3*	*5*	*7*	*9*	*10*	*14*
2	.551	.657	.718	.736	.744	.747	.752
3	.604	.702	.759	.776	.783	.785	.790
5	.680	.766	.813	.827	.833	.835	.839
7	.725	.804	.845	.857	.863	.865	.868
9	.756	.828	.865	.876	.880	.882	.885
10	.769	.839	.874	.885	.889	.890	.893
14	.810	.868	.899	.907	.911	.912	.915

Note: Cell entries are averaged over all item covariances and judgment precision conditions.

14 alternatives. These findings and administrative considerations might suggest that the first alternative would be preferable for telephone interviews and the second for written questionnaires.

Caution should be used, however, in interpreting these findings for two reasons. First, they hold for subject-centered composite scales and any attempt to apply them to other stiuations would be inadvisable. Second, the assumptions underlying the model may not be valid. For example, judgment accuracy and the number of response alternatives were manipulated independently, whereas the conceptual framework suggests that they are related. Thus, contrary to the implications of Table 1, judgment accuracy might be significantly greater for the two items with 14 alternatives than for the nine items with two alternatives. Other significant assumptions involve a uniform distribution of responses and the stability of the error term across the underlying continuum.

As a final point, it should be noted that the results of the simulation differed significantly for the three dependent variables. This fact suggests that the choice of dependent variable should be made judiciously and that the findings based on a single measure of reliability should be interpreted cautiously.

The use of reliability as a criterion in empirical studies to evaluate the number of response alternatives for items in subject-centered scales has a long history. Murphy and Likert (1938, p. 47) suggested that reliability can be improved by increasing the number of items or response alternatives. In one test (1938, p. 55) they found that correlation, as measured by the Spearman-Brown formula, was raised from .88 to .94 by using 36 five-point items instead of 44 three-point items.

Peabody (1962) used data from four different Likert summated attitude scales, each consisting of six-point items, to estimate the sources of variation in respondents' composite scores. He found that between 70 and 80% of the variation was attributable to the direction component of the responses (their sign: + or −), approximately 10% was attributable to the extremeness of the responses (their absolute value: 1, 2, or 3), and between 10 and 20% was attributable to the covariation of the direction and extremeness components. Accordingly, coding the six-point responses dichotomously on the scales consisting of 28, 32, 40, or 208 items involved no great loss in systematic variation in responses.

An even more interesting result was obtained. The direction scores were positively correlated with the extremity of positive responses and negatively correlated with the extremity of negative responses, as one would expect. However, a high positive correlation was found between respondents' use of extreme positive and negative responses on the same attitude scale. Peabody concludes that although the intensity with which individuals respond reflects in part their underlying attitude, it appears to be more reflective of the idiosyncratic response sets of these individuals. These findings provide support for Cronbach's (1950) warning that increasing the number of response alternatives simply because it increases the reliability of a subject-centered scale may only facilitate response sets so as to reduce the scale's validity. If this is the case, perhaps it is fortunate that the extremeness component of responses accounts for such a small portion of the variation in respondents' composite scores.

Komorita (1963) randomly divided a set of attitude statements to form two 14-item scales. Both forms were administered to 286 respondents with the order of presentation varied systematically. Respondents were scored by summing their responses on the six-point items in the conventional manner. Alternatively, the individual items were treated dichotomously and respondents were scored according to the number of items having a favorable response. The correlations between the respondents' two scores were over .95 for both forms, leading to the conclusion that the Likert scale expresses primarily the content of an attitude rather than its intensity.

Komorita also found that the correlations of the respondents' scores for the two forms declined negligibly (from .93 to .91) when the six-point items were treated dichotomously. However, the decline was substantial (from .83 to .71) when the process was repeated for the composite scores based on sets of three items which had been selected at random from the larger scales. This finding suggests that the number of response alternatives has greater impact on reliability when the number of items constituting the scale is small.

In a second study, Komorita and Graham (1965) examined the impact of the number and homogeneity of items on the relationship between reliability and the number of response alternatives. A 24-item "homogeneous" test and a 24-item "heterogeneous" test were identified from previously published evaluations. Four groups of respondents were employed so that each completed one of the composite scales utilizing either two or six response alternatives. As expected, the alpha coefficients for the homogeneous test administrations were substantially greater than those for the heterogeneous test administrations. As hypothesized, the coefficients for the homogeneous test administrations were identical (.92), whereas the coefficient was substantially higher for the six-point administration of the heterogeneous scale than for the two-point administration (.74 versus .62).

In addition, the reliabilities of the scales for 3, 6, 12, and 36 items were estimated by applying the Spearman-Brown formula to the alpha coefficients previously calculated. The results suggest that reliability increases monotonically at a decreasing rate with the number of scale items and that no interaction is present between the number of items and the number of response alternatives.

Jacoby and Matell (1971; also see Matell and Jacoby 1971) administered a scale of personal values consisting of six 12-item subscales to 360 respondents who were assigned to 18 experimental conditions which differed in the number of response alternatives (two through 19) provided by the Likert-type scales. The test was administered twice at a three-week interval. The criteria of test-retest reliability, internal consistency reliability (alpha), concurrent validity, and predictive validity were examined in a one-way analysis of variance.

No significant relationship between either of the measures of reliability or validity and the number of response alternatives was found for any of the subscales. The authors concluded that " . . . reliability should not be a factor in determining a Likert-type scale rating format, because it is independent of the number of scale steps employed" (1971, p. 498). They discuss the flexibility in data collection that this conclusion allows and invite other researchers to investigate whether such flexibility also is possible with other types of scales and other populations.

Jacoby and Matell also collapsed the original data into dichotomous form for the experimental conditions with even-numbered scales and into trichotomous form for the conditions with odd-numbered scales. They found that the resultant reliability and validity measures were slightly lower than those for the data in their original form but that the differences were not statistically significant. The authors conclude that such collapsing of data does not have any "deleterious effects vis à vis reliability and validity" and makes possible the direct comparison of responses obtained by different response formats.

Several points should be considered in evaluating the conclusions drawn by Jacoby and Matell about the optimal number of response alternatives for a Likert-type scale. It is not clear that the distinction between Likert and other types of scales, such as the semantic differential, is pertinent. Likert and semantic differential scale formats may be treated as functionally equivalent (see Kassarjian and Nakanishi 1967; Menezes and Elbert 1979). A more meaningful distinction would seem to be whether a particular scale format was employed in a stimulus-centered or response-centered scaling approach. If so, the generalizations made by the authors are more appropriate for composite subject-centered scales than for Likert-type scales as such. It should also be pointed out that although the conclusions are appropriate for subject-centered scales, they are limited in their external validity for reasons suggested in the preceding discussion. For example, Lehmann and Hulbert (1972) point out that the number of items in the composites studied by Jacoby and Matell mediates the relationship between the number of response alternatives and measures of reliability.

Similarly, caution should be used in interpreting their findings to suggest that dichotomous or trichotomous items are adequate when they are combined in composite scales. Respondents may feel uncomfortable about being forced to be so categorical in their responses. Osgood, Suci, and Tannenbaum (1957, p. 85) report such problems with five-point scales. Rundquist and Sletto (1936, p. 91), Ghiselli (1939), and Matell and Jacoby (1972) found that increasing the number of response alternatives reduced the frequency with which respondents used an "uncertain" or neutral response. Perhaps most significantly, Ghiselli also found that more individuals responded to statements positively when four-point scales were provided than when only two-point scales were used. In several cases, the effect was so great as to reverse the majority opinion from negative to positive.

To summarize, the psychometric literature has provided a fairly clear understanding of the impact of several characteristics of a subject-centered composite scale on its reliability, but has not considered the significance of the number of response alternatives offered by the items comprising the composite. The Monté Carlo research is significant because it introduces the number of response alternatives into the fold of factors that should be considered in scale design, but is necessarily limited by the simulation parameters. The empirical studies provide additional, consistent evidence suggesting that increasing the number of alternatives may increase the reliability of a scale but that the potential is probably minor in comparison with other means.

Looming over these conclusions is Cronbach's warning that an exclusive concern with reliability as a criterion for evaluating the number of response alternatives may simply be encouraging response sets which will have an adverse effect on the validity of the scale. Surprisingly, validity has not been used as a criterion in such studies. The major contribution of Jacoby and Matell is that they did use this criterion and perhaps some assurance can be gained from the fact that they found no significant relationship between either of their validity measures and the number of response alternatives.

CONCLUSIONS

What is apparent from the extensive body of research is that there is no single number of response alternatives for a scale which is appropriate under all circumstances. Though no formula is available to indicate exactly what this number should be for a particular set of circumstances, we have gained a reasonable understanding of some of the important factors which operate along with the number of response alternatives in influencing the quality of information obtained by scales.

If one adopts the analysis of variance perspective presented in the discussion of the conceptual frame-

Table 2
FACTORS INFLUENCING THE QUALITY OF SCALED INFORMATION

Factor	*Stimulus-centered scales*	*Subject-centered scales*
I. Amount of information available for transmission	Homogeneity of stimuli	Homogeneity of respondents
II. Channel capacity of the scale	Number of response alternatives	Number of response alternatives
III. Number of scaling replications	Number of respondents	Number of scale items in the composite
IV. Redundancy among replications	Inter-rater covariation	Interitem covariation
V. Response error	Variation in responses not explained by stimulus differences	Variation in responses not explained by respondent differences

work, some general statements can be made about these factors as they influence both stimulus-centered and subject-centered scales. Five of these factors are presented in Table 2.

The first of the factors is the amount of information available for transmission by the scale. This information may be viewed as the variation in stimuli in the case of stimulus-centered scales or in respondents in the case of subject-centered scales. If the stimuli or respondents in a set are identical then no scaling is required, and the potential for success of scaling efforts under other circumstances is constrained by the homogeneity of the objects (stimuli or respondents) being scaled. Because the choice of these objects is not usually determined as a part of the research design, this is the only one of the five factors which is beyond the control of the researcher.

The second factor is the channel capacity of the individual scale item. This is the only factor about which there is any certainty, for we know that the channel capacity of an individual scale increases monotonically with the number of response alternatives. Though the channel capacity in metric terms is limited significantly for two or three alternatives, diminishing returns are achieved by adding more alternatives. Effectiveness aside, this is the factor most easily manipulated by the researcher.

The number of scaling replications is the third factor. In the case of subject-centered scales, redundant information is obtained by adding more items to the composite scale, whereas such information is acquired by adding more judges (i.e., respondents) for stimulus-centered scales. It should be emphasized that the "redundancy" of information is a positive quality suggesting that the replications provide the supportive or confirming information vital to many psychometric techniques and should not be confused with the popular connotation of the term which suggests something to the contrary.

The fourth factor is the extent of redundancy among the replications. Though the amount of redundant information can be increased by adding more replications, it can also be increased by selecting only those replications which can offer a large degree of redundant information in the first place. Researchers increase the amount of redundant information in a subject-centered composite scale by selecting those items with high covariances. It is interesting to note that although the rater reliability of stimulus-centered scales can in the same manner be increased by selecting individuals whose responses have high covariances, this is not done in marketing as respondents are virtually always selected on the basis of statistical criteria rather than measurement criteria.

The fifth factor influencing the quality of information obtained by scales is response error. This is the factor we know least about. Response error is no doubt closely related to the amount of information among the stimuli available for transmission, the channel capacity of the individual scale, the extent of redundancy among the replications, as well as the ability and interest of respondents, but the relationships among these factors are not known and much additional research is needed in this area.

Response error can also be influenced by the physical presentation of the scales. For example, the quantity and quality of labeling for the alternatives have been found to be important. Rundquist and Sletto (1936, p. 89ff) and Katz (1944) report that responses on graphic scales tend to concentrate around the positions of the scale that have been defined.

Areas in Need of Additional Research

Although much replication and extension is needed generally, two areas are in particular need of additional research. First, much more needs to be known about reponse error and response bias in the context of scales. Surprisingly little is known about the processes of psychological judgment. Though some of the psychophysical methods may be equally applicable to physical and psychological stimuli, it is not clear that the actual processes are the same in both cases. If, as Volkmann (1951) has suggested, there is the basis for a "general psychology of discrimination," significant advances in our understanding of scales can be made rather quickly. However, it is not yet clear what general principles exist.

The second and probably more immediate need for research is in the development of methods which can be used effectively in the pretesting stage of research to evaluate the impact of the number of response alternatives on the quality of the information being collected. This need is particularly great in the case of stimulus-centered scales. Such methods might be based on a split-balloting scheme or might be more analytical and involve the direct assessment of the homogeneity of the objects being scaled, for example.

Recommendations for the Applied Researcher

Despite the trepidation experienced in attempting to make specific recommendations based on this review, the author would not support a moratorium on the use of scales until more definitive research has been conducted. Certain things do appear clear. First, scales with two or three response alternatives are generally inadequate in that they are incapable of transmitting very much information and they tend to frustrate and stifle respondents. Second, the marginal returns from using more than nine response alternatives are minimal and efforts for improving the measurement instrument should be directed toward more productive areas. Third, an odd rather than an even number of response alternatives is preferable under circumstances in which the respondent can legitimately adopt a neutral position. Overuse of the neutral category by respondents can generally be avoided by providing them with an adequate number of reasonable response alternatives. Fourth, even a few response alternatives may be too many for the respondent if comprehensible instructions and labeling of response alternatives are not included to enable the respondent to conceptualize and respond in spatial terms.

It is ironic that the magic number seven plus or minus two appears to be a reasonable range for the optimal number of response alternatives, despite the fact that Miller's review is not directly relevant to this question. Several factors seem to suggest the number of alternatives within this range of five to nine which would be appropriate under a particular set of circumstances. In the case of subject-centered scales, five alternatives seem adequate for the individual items and energy is best spent on increasing the number of quality items constituting the composite scale. In the case of stimulus-centered scales, as many as nine alternatives may be appropriate if the stimuli are heterogeneous and the respondents are sophisticated with regard to these stimuli and are committed to their scaling task.

REFERENCES

Attneave, Fred (1949), "A Method of Graded Dichotomies for the Scaling of Judgments," *Psychological Review*, 56 (November), 334-40.

——— (1959), *Applications of Information Theory to Psychology*. New York: Holt, Rinehart and Winston.

Bartlett, J. C. and Calvin G. Green (1966), "Clinical Prediction: Does One Sometimes Know Too Much?" *Journal of Counseling Psychology*, 13 (Fall), 267-70.

Behar, Isaac (1963), "On the Relation Between Response Uncertainty and Prediction Time in Category Judgments," *Perceptual and Motor Skills*, 16 (April), 595-6.

Bendig, A. W. (1953), "The Reliability of Self-Ratings as a Function of the Amount of Verbal Anchoring and of the Number of Categories on a Scale," *Journal of Applied Psychology*, 37 (February), 38-41.

——— (1954a), "Reliability and the Number of Rating Categories," *Journal of Applied Psychology*, 38 (February), 38-40.

——— (1954b), "Transmitted Information and the Length of Rating Scales," *Journal of Experimental Psychology*, 47 (May), 303-8.

——— and J. B. Hughes (1953), "Effect of Amount of Verbal Anchoring and Number of Rating-Scale Categories Upon Transmitted Information," *Journal of Experimental Psychology*, 46 (August), 87-90.

Benson, Purnell H. (1971), "How Many Scales and How Many Categories Shall We Use in Consumer Research?—A Comment," *Journal of Marketing*, 35 (October), 59-61.

Best, Roger J., Gerald S. Albaum, Del I. Hawkins, and Georgia Kenyon (1978), "Number of Response Intervals and Reliability of Factor Analyzed Semantic Scale Data," Southwestern Marketing Association.

Bevan, William and Lloyd L. Avant (1968), "Response Latency, Response Uncertainty, Information Transmitted and the Number of Available Judgmental Categoiies," *Journal of Experimental Psychology*, 76 (March), 394-7.

Boyce, Arthur C. (1915), "Methods for Measuring Teachers' Efficiency," in *The Fourteenth Yearbook of the National Society for the Study of Education*, Part II, S. Chester Parker, ed. Chicago: University of Chicago Press.

Bricker, P. D. (1955), "The Identification of Redundant Stimulus Patterns," *Journal of Experimental Psychology* 49 (February), 73-81.

Carroll, John B. (1961), "The Nature of Data, or How to Choose a Correlation Coefficient," *Psychometrika*, 26 (December), 347-72.

Carroll, J. D. (1972), "Individual Differences and Multidimensional Scaling," in *Multidimensional Scaling: Theory and Application in the Behavioral Sciences*, Volume 1, Roger N. Shepard et al., eds. New York: Seminar Press.

Champney, Horace and Helen Marshall (1939), "Optimal Refinement of the Rating Scale," *Journal of Applied Psychology*, 23 (June), 323-31.

Conklin, Edmund S. (1923), "The Scale Values Method for Studies in Genetic Psychology," *University of Oregon Publication*, 2, 3-36.

Connor, Robert J. (1972), "Grouping for Tests of Trend in Categorical Data," *Journal of the American Statistical Association*, 67 (September), 601-4.

Coombs, Clyde H. (1964), *A Theory of Data*. New York: John Wiley & Sons, Inc.

Cronbach, Lee J. (1950), "Further Evidence on Response Sets and Test Design," *Educational and Psychological Measurement*, 10 (Spring), 3-31.

——— (1951), "Coefficient Alpha and the Internal Structure of Tests," *Psychometrika*, 16 (September), 297-334.

Ebel, Robert L. (1951), "Estimation of the Reliability of Ratings," *Psychometrika*, 16 (December), 407-24.

Garner, W. R. (1960), "Rating Scales, Discriminability, and

Information Transmission," *Psychological Review*, 67 (November), 343-52.

——— and H. W. Hake (1951), "The Amount of Information in Absolute Judgments," *Psychological Review*, 58 (November), 446-59.

Ghiselli, Edwin E. (1939), "All or None Versus Graded Response Questionnaires," *Journal of Applied Psychology*, 23 (June), 405-15.

——— (1964), *The Theory of Psychological Measurement*. New York: McGraw-Hill Book Company.

Green, Paul E. and Vithala R. Rao (1970), "Rating Scales and Information Recovery—How Many Scales and Response Categories to Use?" *Journal of Marketing*, 34 (July), 33-9.

Guest, Lester (1962), "A Comparison of Two-Choice and Four-Choice Questions," *Journal of Advertising Research*, 2 (March), 32-4.

Guilford, J. P. (1938), "The Computation of Psychological Values from Judgments in Absolute Categories," *Journal of Experimental Psychology*, 22 (January), 32-42.

——— (1954), *Psychometric Methods*, second edition. New York: McGraw-Hill Book Company.

Guttman, Louis (1950), Chapters 2, 3, 6, 8, and 9 in *Measure and Prediction*, Samuel A. Stouffer et al., eds. Princeton, New Jersey: Princeton University Press.

Hulbert, James (1975), "Information Processing Capacity and Attitude Measurement," *Journal of Marketing Research*, 12 (February), 104-6.

Jacoby, Jacob and Michael S. Matell (1971), "Three-Point Scales are Good Enough," *Journal of Marketing Research*, 8 (November), 495-500.

Jenkins, G. Douglas, Jr. and Thomas D. Taber (1977), "A Monte Carlo Study of Factors Affecting Three Indices of Composite Scale Reliability," *Journal of Applied Psychology*, 62 (August), 392-8.

Jones, Richard R. (1968), "Differences in Response Consistency and Subjects' Preferences for Three Personality Inventory Response Formats," *Proceedings of the 76th Annual Convention of the American Psychological Association*, 247-8.

Kassarjian, Harold H. and Masao Nakanishi (1967), "A Study of Selected Opinion Measurement Techniques," *Journal of Marketing Research*, 4 (May), 148-53.

Katz, Daniel (1944), "The Measurement of Intensity," in *Gauging Public Opinion*, Hadley Cantril, ed. Princeton, New Jersey: Princeton University Press, 51-65.

Kelley, Truman L. (1923), *Statistical Method*. New York: Macmillan.

Komorita, S. S. (1963), "Attitude Content, Intensity, and the Neutral Point on a Likert Scale," *Journal of Social Psychology*, 61 (December), 327-34.

——— and William K. Graham (1965), "Number of Scale Points and the Reliability of Scales," *Educational and Psychological Measurement*, 25 (November), 987-95.

Lehmann, Donald R. and James Hulbert (1972), "Are Three-Point Scales Always Good Enough?" *Journal of Marketing Research*, 9 (November), 444-6.

Likert, Rensis (1932), "A Technique for the Measurement of Attitude," *Archives of Psychology*, no. 140, 1-55.

Lissitz, Robert W. and Samuel B. Green (1975), "Effect of the Number of Scale Points on Reliability: A Monte Carlo Approach," *Journal of Applied Psychology*, 60 (February), 10-3.

Lunney, Gerald H. (1970), "Using Analysis of Variance with a Dichotomous Dependent Variable: An Empirical Study," *Journal of Educational Measurement*, 7 (Winter), 263-9.

Martin, Warren S. (1973) "The Effects of Scaling on the Correlation Coefficient: A Test of Validity," *Journal of Marketing Research*, 10 (August), 316-8.

——— (1978), "Effects of Scaling on the Correlation Coefficient: Additional Considerations," *Journal of Marketing Research*, 15 (May), 304-8.

———, Benjamin Fruechter, and William J. Mathis (1974), "An Investigation of the Effects of the Number of Scale Intervals on Principal Components Factor Analysis," *Educational and Psychological Measurement*, 34 (Autumn), 537-45.

Matell, Michael S. and Jacob Jacoby (1971), "Is There an Optimal Number of Alternatives for Likert Scale Items? Study I: Reliability and Validity," *Educational and Psychological Measurement*, 31 (Autumn), 657-74.

——— and ——— (1972), "Is There an Optimal Number of Alternatives for Likert Scale Items?" *Journal of Applied Psychology*, 56 (December), 506-9.

McGill, William J. (1954), "Multivariate Information Transmission," *Psychometrika*, 19 (June), 97-116.

Menezes, Dennis and Norbert F. Elbert (1979), "Alternative Semantic Scaling Formats for Measuring Store Image: An Evaluation," *Journal of Marketing Research*, 16 (February), 80-7.

Miller, George A. (1955), "A Note on the Bias of Information Estimates," in *Information Theory in Psychology: Problems and Methods*, Henry Quastler, ed. Glencoe, Illinois: The Free Press, 95-100.

——— (1956), "The Magical Number Seven, Plus or Minus Two: Some Limits on Our Capacity for Processing Information," *Psychological Review*, 63 (March), 81-97.

——— and Frederick C. Frick (1949), "Statistical Behavioristics and Sequences of Responses," *Psychological Review*, 56 (November), 311-24.

Morrison, Donald G. (1972), "Regression with Discrete Dependent Variables: The Effect on R^2," *Journal of Marketing Research*, 9 (August), 338-40.

Murphy, Gardner and Rensis Likert (1938), *Public Opinion and the Individual*. New York: Harper and Brothers Publishers.

Nishisato, Shizuhiko and Yukhiko Torii (1971), "Effects of Categorizing Continuous Normal Variables," *Japanese Psychological Research*, 13 (May), 45-9.

Osgood, Charles E., George J. Suci, and Percy Tannenbaum (1957), *The Measurement of Meaning*. Chicago: University of Chicago Press.

Peabody, Dean (1962), "Two Components in Bipolar Scales: Direction and Extremeness," *Psychological Review*, 69 (March), 65-73.

Pemberton, H. Earl (1933), "A Technique for Determining the Optimal Rating Scale for Opinion Measures," *Sociology and Social Research*, 17 (May-June), 470-2.

Peter, J. Paul (1979), "Reliability: A Review of Psychometric Basics and Recent Marketing Practices," *Journal of Marketing Research*, 16 (February), 6-17.

Peterson, Robert A. and Subhash Sharma (1977a), "A Note on the Information Content of Rating Scales," *Proceedings of the American Marketing Association*, 324-6.

——— and ——— (1977b), "Adjusting Correlation Coefficients for the Effects of Scaling," *Proceedings of the American Statistical Association*.

Pollack, Irwin (1952), "The Information of Elementary Auditory Displays," *Journal of the Acoustical Society* *of America*, 24 (November), 745-9.

Ramsay, J. O. (1973), "The Effect of Number of Categories in Rating Scales on Precision of Estimation of Scale Values," *Psychometrika*, 37 (December), 513-32.

Rundquist, Edward A. and Raymond F. Sletto (1936), *Personality in the Depression*. Minneapolis: The University of Minnesota Press.

Saffir, Milton (1937), "A Comparative Study of Scales Constructed by Three Psychophysical Methods," *Psychometrika*, 2 (September), 179-98.

Shannon, Claude and Warren Weaver (1949), *The Mathematical Theory of Communication*. Urbana: University of Illinois Press.

Sheppard, W. F. (1898), "On the Calculation of the Most Probable Values of Frequency Constants for Data Arranged According to Equi-Distant Divisions of the Scale," *Proceedings of the London Mathematical Society*, 29 (March), 353-80.

Stevens, S. S. and Hallowell Davis (1938), *Hearing*. New York: John Wiley & Sons, Inc.

Symonds, Percival M. (1924), "On the Loss of Reliability in Ratings Due to Coarseness of the Scale," *Journal of Experimental Psychology*, 7 (December), 456-61.

Thurstone, Louis Leon (1928), "Attitudes Can Be Measured," *American Journal of Sociology*, 33 (January), 529-54.

——— and E. J. Chave (1929), *The Measurement of Attitude*. Chicago: University of Chicago Press.

Torgerson, Warren J. (1958), *Theory and Methods of Scaling*. New York: John Wiley & Sons, Inc.

Volkmann, John (1951), "Scales of Judgment and Their Implications for Sociology," in *Social Psychology at the Crossroads*, J. H. Rohrer and M. Sheriff, eds. New York: Harper and Brothers, 272-94.

Young, Forrest W. (1970), "Nonmetric Multidimensional Scaling: Recovery of Metric Information," *Psychometrika*, 35 (December), 455-73.