PURPOSES OF SCALING TECHNIQUES AND THE CHOICE AMONG THEM

Robert P. Abelson, Yale University

Scaling can basically be defined as the establishment of rules by which a set of numbers can be assigned to a set of magnitudes of a property. Or, more simply, one can talk of assigning sets of numbers to set of objects which embody different magnitudes. In psychology, the objects may be people, physical stimuli, attitude statements, objects of preference, such as aesthetic objects or commodities, and so on.

A psychological scale using people as objects might be a scale of intelligence, a scale of some personality trait, or a scale of position on a controversial social issue; using stimuli, one might generate scales of subjective brightness loudness, heaviness, etc; a scale of attitude statements might specify the degree of pro-ness or con-ness of each; a scale of object preference would express the perceived value or utility of each object. There are other varieties of examples which could be given, but one thing is quite clear without proliferating examples; psychological scaling is at the outset a risky and speculative venture. The properties of intelligence, personality attributes, attitude, value, and even sensation are not well enough understood for quantification to be an easy and straightforward matter. One hopes, of course, that the attempt to scale a given psychological property will lead to some understanding of the property. But understanding in turn is often pre-requisite to making a reasonable try at scaling. It is a circular affair, like an inexperienced actor trying to break into show business.

Urgent need, however, has inspired the development of a variety of procedures for psychological scaling. I will discuss some of them briefly, without attempting to give a systematized account of the entire sweep of methods. This has been done very excellently in the recent book by Warren Torgerson (1958).

First of all, let me exclude from present consideration those methods which attempt to assign numbers to attributes of persons. There is in principle a duality between scales of persons derived by having persons respond to stimuli (e.g., test items) and scales of stimuli derived by having them judged or responded to by persons. (Mosier, 1940; Coombs, 1956) In practice, however, there are qualitative differences between scales of people and scales of stimuli. People are more complex than stimuli; it is harder to conceptualize properties of people than properties of stimuli. People are more changeable than stimuli; next year's scale of people is apt to be different from today's. People have less conveniently available time than stimuli; it is easy to use the same stimuli again and again to learn more of their properties, but not so easy to use the same people again and again. In addition, there is a curious sampling conundrum upsetting the duality -- if you want to make a stimulus scale more reliable, you can simply use more people to respond to or judge the stimuli, sampling these people from a specified population; but if you want to make a person scale more reliable, you often cannot sample more stimuli for people to respond to -- you must create or select new stimuli from an unspecified population.

The upshot of these considerations (and others) is that scales of stimuli should be more stable than scales of people, as apparently they are. I discuss all this because it will become important later.

In discussing specific scaling procedures, I will want to distinguish between those techniques which make assumptions about the psychological nature of the responses to the stimuli and those which merely seek useful quantitative scales without establishing a formal model.

Certain specific scaling procedures

Paired Comparisons

Experimentally, paired comparisons involves presenting successive pairs of stimuli to subjects who judge which member of each pair is darker, tastier, more preferred, or whatnot.

The method known to psychologists for analysing such data comes from Thurstone's Law of Comparative Judgment. (Thurstone, 1927) Briefly, this law (or better, hypothesis) states that any stimulus of judgment or choice activates response processes which are to some extent variable, between individuals or within individuals; further, that the distribution of potential responses in Gaussian. The probability that stimulus A will be chosen over stimulus B depends upon the means and standard deviations of the A-distribution and the B-distribution and upon the assumed correlation between responses when A and B are presented concurrently. The scale value of each stimulus is defined as the mean of its response distribution. The matrix of proportions of times each stimulus is chosen over every other is sufficient to determine within an arbitrary linear transformation the scale values of the stimuli. The normal distribution assumption is necessary to get the analysis off the ground, although other distribution forms are equally plausible. In fact, the paired comparison method developed by Bradley and Terry (1952) though couched in somewhat different terms, can in part be viewed as the equivalent of the Thurstone model, except that the logistic distribution is used instead of the normal distribution (Gridgeman, 1955).

These two methods, then, make some claim to an understanding of a reality underlying the choices. These claims lead to internal consistency checks on the methods, so that applications yield more than scales; they provide tests of the models (Bradley, 1954, 1955; Mosteller, 1951). Unfortunately these tests happen not to be very sensitive to minor departures.

One other paired comparison method, due to Scheffè (1952) requires that the subject not only choose a member of each pair, but also specify the degree to which the member is the more preferred. Scheffè's method also makes assumptions, but these are in the spirit of statistical assumptions rather than psychological assumptions. Other methods, employing ranks or rankits, or the ingenious manipulations of paired comparisons matrices suggested by Kendall (1955), usually make no pretense to being formal models of the judgment process.

Categorical or single stimulus methods

The single stimulus methods present subjects with one stimulus at a time and require a numerical rating of the stimulus. The placements can be taken as they stand, or else some model can be imposed on the data.

A popular model in psychology is the so-called successive intervals model (Edwards and Thurstone, 1952; Diederich, Messick, and Tucker, 1957). This derives again from Thurstone's judgment model. The boundaries between adjacent ordered categories are assigned numerical values on the same underlying continuum as the Gaussian response distributions. There are internal consistency checks on the method, as with the Thurstone paired comparisons model.

Guttman scales

Guttman scaling (Guttman, 1941, 1947) is primarily an assumption about the nature of reality rather than a scaling technique per se. A set of stimuli or items are presented to a set of individuals, who respond either positively or negatively to each item (say, the item might be a statement and the individual is to agree or disagree with it). It is assumed that there exists an ordering of the items such that if an individual agrees with a given item he also agrees with all items beneath the given one in the ordering. Thus the entire set of items constitutes a kind of staircase, with individuals grouped according to which level along the staircase they occupy. Empirically it always turns out that the staircase pattern is not perfect; an index called the Reproducibility of the scale specifies the proportion of responses which conform to the perfect pattern, and investigators are usually happy when this index reaches .90. It has become a social science parlor game to see what new types of responses achieve scales with acceptable reproducibilities. The chief criticism of this game is that as a model of the nature of reality, the Guttman scale is rarely convincing -- the Reproducibilities generally lie about half-way between chance level and the perfection required by the model -- and when the requirement of perfection is far from being met, then the method loses its special appeal as a means of obtaining clean scales. It becomes no less arbitrary a procedure than the next man's.

Magnitude estimation

It has been argued by S. S. Stevens and others (Stevens, 1957; Stevens and Galenter, 1957) that scaling methods based upon the use of variability of response, like paired comparisons and successive intervals, are misguided. The use of intra-subject confusion on which to erect the edifice of measurement is decried. If you want estimates of subjective magnitudes, says Stevens, ask the subject directly. Thus to scale loudnesses, for example, one might give the subject a preliminary tone and tell him to call its loudness 10. Then the subject attaches numerical values to subsequent stimuli in accordance with his subjective estimate of their loudness. A related device in the field of consumer preference would be to ask the subject how many dollars he would be willing to pay for each of a number of items.

The magnitude estimation methods have been making considerable recent headway in psychophysics. It appears that on a large number of physical continua such as loudness, brightness, heaviness, etc., there is a common function relating subjective magnitude to objective magnitude. The continued use of the method depends heavily upon the discovery of such broad empirical relationships, for there is no internal consistency check on the method to verify its validity as a model. Indeed it is not intended to be model at all, but rather a straight-forward means for generating numerical scales.

Other methods

There are a variety of scaling methods devised by Coombs (1950, 1952, 1954, 1958) which are similar to Guttman scaling in that perfection of response pattern is required, but which by and large are more flexible. In particular, they allow for multi-dimensional scales, and offer the possibility of scales of objects as viewed by single judges, rather than sets of judges. The level of measurement achieved is not numerical, but something short of this -- so-called ordered metric measurement.

Another technique resulting in ordered metric scales uses a lottery to establish relative scale positions of objects (Siegel, 1956). Subjects have been asked such piquant questions as, "Would you rather: a) admit Negroes to your school for sure, or b) elect a 50-50 bet, with heads signifying that one of your friends marries a Negro and tails that Negroes be allowed to ride in local buses"? A subject selecting alternative a) is inferred to hold a greater value or utility difference between intermarriage and school desegregation than between school desegregation and bus desegregation.

Further new techniques may soon flow from a model of choice behavior conceived by Duncan Luce (1959). The basic axiom of this model is essentially that when one object is to be chosen from a set of objects, the ratio of the probability of choice of A vs. B is independent of the number and nature of other objects presented along with A and B for choice.

Studies comparing scaling methods

Though not all methods have been compared with all other methods, a number of comparison studies have been made. Mosteller (1958) compared a variety of assumptions for the underlying distribution in the paired comparisons model -- rectangular, arc sine, normal, exponential, and a t-distribution with high tails -- and found that it makes almost no difference which one is used. The correlations between the resulting scales run from .9965 to .99998. The scales may be characterized as very weak quadratic transformations of one another.

Jackson and Fleckstein (1957) compared a variety of experimental and analytic methods involving paired comparisons. The results were extraordinarily insensitive to choice of method. A representative correlation is .975. Bliss, Greenwood and White (1956) made even more extensive comparisons of methods, including ranking techniques as well as paired comparisons. Again the picture is one of fantastically high correlations among outcomes, especially when differing analytical methods are used on the same experimental data. When different types of experimental data are compared, as with Scheffé's technique vs. a version of Kendall's technique the correlation between scales drops to its lowest, a rock bottom value of .990. Jones (1958), Gulliksen (1953), and others have compared different analytical solutions for successive interval data. The correlations are in the .99 range. Kelley, Hovland et. al.(1955) found a correlation of .910 between paired comparison and successive intervals values. This even in a situation where the category assignments required for the successive intervals method were very deviantly performed by the judging group.

It is not surprising that paired comparisons and successive intervals scales are very closely related, for after all, they are both based on Thurstone's judgment model. Nor is it surprising that paired comparisons scales correlate highly with scales derived from rank ordering of the stimuli. (Incidentally, these correlations can be astronomical. In one study by Ross (1955), these two methods correlated .998). Transitivity of choice (in a statistical, rather than an absolute sense) is an unexceptioned finding in investigations to date (Davidson & Marschak, 1957; Davis, 1958). No matter whether rank ordering of objects is done by separate pairs or all at once, the results are essentially the same. There is good reason to expect, too, that paired comparisons scales extracted by Thurstone's method or the Bradley-Terry method will prove almost identical. Thus many of the high correlations between scales derived by different methods can be rationalized as mathematical necessities or as natural consequences of mundane empirical regularities.

However, there are some correlations between scales which are not so readily explained. Three disparate methods have been applied to scale a set of nine color chips multi-dimensionally -- that is, where scale values on more than one dimension are assigned to each stimulus. The three methods were: a successive intervals method extended to the multidimensional case (Messick, 1956) the so-called complete method of triads due to Torgerson (1952) and Shepard's method (1957) which uses probabilities of confusion in learning responses to the stimuli. Shepard (1958) has given the correlations between the three methods on two dimensions as around .98. Other striking examples of high correlation using disparate methods occur in psychophysics. A most interesting example is a recent study by Galanter and Messick (1958). They applied both successive intervals scaling and the direct magnitude estimation method to a set of stimuli varying in physical loudness. The psychological loudness determined by the two methods correlate .896. I will say more about this study later. A study by Benson and Platten (1956) comparing these same two methods in the domain of preferences yields a correlation of .968.

There are other studies which could be cited, but the long and short of the situation is that it is almost impossible to find two scaling methods which when applied to the same stimuli will yield anything less than a very high correlation. It appears that if the objects of a given psychological domain possess sufficient underlying order so that some one scaling method is reliable, then any other reasonable scaling method will reveal essentially the same order. Any knife that cuts at all, even the bluntest, will expose the same corpus. Mind you, I am talking about scales of stimuli, not about scales of people. Method comparisons on the latter yield correlations that run eighty-ish down through fifty-ish or lower. But as I have already indicated, there are several reasons for expecting a certain instability in people-scales.

What is the nature of the underlying stimulus constancies upon which different methods converge? Is it that a stimulus scale basically amounts to a rank order and that all methods reveal the true rank order? No, for even with a rank order correlation of unity among n objects, the product-moment correlation can be as low as $1/\sqrt{n-1}$, or .33 for ten objects, .20 for twenty-six objects. Even if one deals not with minimum possible correlations but with expected or average correlations in some sense, it is still evident that the obtained correlations in comparisons of scaling methods are much higher than could be accounted for by postulating shared rank order alone. Something even stronger than mere ordinality must typically underly the set of stimuli. There are in principle many gradations between ordinal measurement and cardinal measurement. A very interesting pursuit would be to try to find the level of measurement coordinate with the magnitude of correlations commonly found between scaling methods. Recent Work by Tukey and myself may conceivably shed some light on this question (Abelson & Tukey, 1958).

The objection may be raised that product-moment correlation is not the most appropriate indication of degree of equivalence between two scales. A very high correlation occurs when one scale is but a mild transformation of the other, yet the distortion might be the thing which most interested us. A very high correlation would also occur if two scales were linearly related save for a small number of deviant stimuli displaced from the line of realtionship. This displacement itself might be the phenomenon of interest. Examples exist both of mild transformation and of idiosyncratic discrepancies or "bumps".

In order to know whether scaling method makes a difference, we must inquire into the investigator's purposes. For some purposes, the choice of a particular scaling technique may not matter at all, while for others it may.

In order to draw some kind of a sample of purposes from the population of current purposes, I have inspected most of the scaling articles in the psychological literature, 1954-1957, listed in a previous comprehensive review (Messick & Abelson, 1957). For each published study, I noted the investigator's purpose, and asked whether the conclusions would have been different had he chosen some other scaling method. I assumed that the worst that could have happened via change in method would have been the introduction of a few slight bumps in the scale or else a mild transformation of the entire scale. I classified the answers to the question "Would it have made a difference?" into three categories: yes, no, and maybe. In the abstract for this paper, the figure 10% is given for the percent of "yes" answers. At this juncture I feel that this percentage, though not incorrect, is superficial. It takes no account of the relative importance of various studies. I think my best course at this point is to give examples of studies varying in the degree to which the scaling method made a difference.

At one extreme are studies in which the purpose is to verify the existence of a gross effect upon judgments of experimental or natural conditions. For instance, it has been demonstrated that when two foods are tasted at one time point and preference judgments elicited at a later time point, the difference in degree of preference is far less than if the preference judgments are given at the time of tasting (Schwartz & Pratt, 1956). Any method of preference scaling would doubtless show the same effect. A study comparing the basic values of college students in the U. S., India, and other countries showed huge differences in value-orientation (Morris & Jones, 1955). Any scaling method would have found the essence of these differences. And so on. Of course, where experimental conditions differ not grossly, but subtly, there is more chance for scaling method to make a difference. However, suppose that two scaling methods differ only in that one is a mildly non-linear transformation of the other. Common experience with the analysis of variance would indicate that sharp differences in outcome are not at all likely to occur under these circumstances (though, to be sure, the analysis may feel more satisfying in one version than in another). The choice between methods is then based upon experimental convenience, richness of by-product information, and the aesthetics of the analysis.

At the other extreme are studies where the scaling method may make a crucial difference. I will cite four of these, typifying the kind of purposes involved.

The previously cited study by Kelley et. al. had both Negro judges and white judges scale the favorableness of statements about Negroes via paired comparisons. It turned out that the scales produced by the two sets of judges were almost perfectly linearly related except for three statements which in the white scale were displaced slightly but discernably above where they were in the Negro scale. Successive interval scaling did not reveal this effect. The displacement happens to be convincing in this case because the three displaced statements concerned "separate but equal" treatment of Negroes. The Negro judges did not consider these so favorable to Negroes, relatively, as did whites.

Another kind of crucial dependence upon scaling method occurs when the investigator wants to know the functional dependence of a psychological scale upon a physical or natural scale. Davidson, Suppes, and Siegel (1957) scaled the utility of very small amounts of money by a lottery method similar to the one mentioned before. With most subjects, utility turned out to be very nearly linear in money. Had another variety of scaling method been used, utility might have turned out, say, a square root function of monetary amount.

In the study by Galanter and Messick mentioned earlier, a successive intervals scale of loudness yielded a rather different function for the relation between subjective and objective loudness than did the magnitude estimation method (even though the two methods correlated .896).

In these two examples, scaling method makes a difference to the extent that the resulting function makes a difference. But here we encounter a new consideration. The successive intervals method yields a logarithmic psychophysical law; the magnitude method yields a power law. One may go from the latter to the former by a logarithmic transformation and come back with an exponential transformation. The two methods are equivalent in the larger sense that it is possible to go from one to the other Who is to say which method is "better"? The issue seems again to come down to a question of convenience.

One further example may clarify the picture. In a study by Cliff (1956), evaluative adjective, and adjective-adverb combinations were scaled for intensity by the successive intervals method. Cliff then sought the relationship between the judged favorableness of the combinations, such as "rather evil", "decidedly charming", etc. and the individual components. He concludes that adverbs exert a multiplicative effect upon adjectives. The quantitative support for this conclusion is really remarkably good. Let us imagine the magnitude methods applied to the same stimuli yielding scale values exponentially related to Cliff's values. His law of adverbs would then be lost. One would need the extra insight, "Take the logarithm of the scale values and then there will be a multiplicative law". The choice between scaling methods yielding different functions only begins to matter when these functions are imbedded in still further operations.

One may also stop and ask, "How is it that one method comes to yield a scale that is, say, logarithmically related to another? What is going on in the judgment process anyhow? This question can be asked of the successive intervals and magnitude methods and is as yet unanswered to everybody's satisfaction. One needs a superordinate model of the judgment process which will explain both types of scale outcomes. Thus, as has happened in other areas of psychology, one starts out with an interest in psychological content matter, constructs measurement devices to deal with that content, and ends up studying the interaction between measurement and measuree instead of the original content matter. Explicit models of judgment or choice processes, like Luce's, thereby assume great importance.

Summary

In most applications of psychological scaling techniques to objects other than people, the results are monumentally indifferent to the choice of scaling method. For most purposes, the choice of method should be made in terms of convenience. Three exceptions to this generalization are: 1) When the conclusions depend upon slight bends or bumps along the scale of stimuli, 2) When a functional relationship is sought between the stimulus scale and some external scale, provided that the function enters into some further consequences, 3) When the interest is not really in scale values at all but rather in a model of what is going on in the response, judgment, or preference process.

When looking for bumps or bends, one wants to choose the method that reveals meaningful bumps clearly and at the same time suppresses false or phrenological bumps. When looking for functional relationships, one wants to choose the scale methods yielding the neatest relationships. When testing models, one of course uses methods appropriate to test the particular model at hand.

It is extremely doubtful that any single method will prove superior on all counts. We will probably have to continue to live with too many methods, differences that make no difference, and too few models.

References

Abelson, R. P. and Tukey, J. W. (1958) Efficient utilization of non-numerical information in quantitative analysis. (In preparation).

Benson, P. H. and Platten, J. H. (1956) Preference measurement by the methods of successive intervals and monetary estimates. J. appl. Psychol., 40, 412-414.

Bliss, C. I., Greenwood, Mary L., and White, Edna S. (1956) A rankit analysis of paired comparisons for measuring the effect of sprays on flavor. Biometrics, 12, 381-403.

Bradley, R. A., and Terry, M. E. (1952) The rank analysis of incomplete block designs. I. The method of paired comparisons. Biometrika, 39, 324-345.

Bradley, R. A. (1954) Incomplete block rank analysis: On the appropriateness of the model for a method of paired comparisons. Biometrics, 10, 375-390.

Bradley, R. A. (1955) Rank analysis of incomplete block designs III. Some large-sample results on estimation and power for a method of paired comparisons. Biometrika, 42, 450-470

Cliff, N. (1956) The relation of adverb-adjective combinations to their components. ONR Technical Report. Princeton, N. J.: Princeton University.

Coombs, C. H. (1950) Psychological scaling without a unit of measurement. Psychol. Rev., 57, 145-158.

Coombs, C. H. (1952) A theory of psychological scaling. Ann Arbor: University of Michigan Press. Eng. Res. Bull. no. 34.

Coombs, C. H. (1954) A method for the study of interstimulus similarity. Psychometrika, 19, 183-194.

Coombs, C. H. (1956) The scale grid: some interrelations of data models. Psychometrika, 21, 313-330.

Coombs, C. H. (1958) An application of a nonmetric model for multidimensional analysis of similarities. Psychol. Reports. (In press).

Davidson, D., and Marschak, J. (1957) Experimental tests of stochastic decision theory. Cowles Foundation Discussion Paper #22. New Haven, Conn.: Yale Univ. (Mimeo).

Davidson, D., Suppes, P., and Siegel, S. (1957) Decision making: An experimental approach. Stanford, Calif.: Stanford Univ. Press.

Davis, J. M. (1958) The transitivity of preferences. Behavioral Sci., 3, 26-33.

Diederich, Gertrude, Messick, S. and Tucker, L. (1957) A general least squares solution for successive intervals. Psychometrika, 22, 159-173.

Edwards, A. L. and Thurstone, L. L. (1952) An internal consistency check for scale values by the method of successive intervals. Psychometrika, 17, 169-180.

Galanter, E. and Messick, S. (1958) The relation between category and magnitude scales of loudness. (In preparation).

Gridgeman, N. T. (1955) The Bradley-Terry probability model and preference tasting. Biometrics, 11, 335-343.

Gulliksen, H. (1954) A least squares solution for successive intervals assuming unequal standard deviations. Psychometrika, 21, 125-134.

Guttman, L. (1941) The quantification of a class of attributes: A theory and method of scale construction. In Horst, Paul, et. al., The Prediction of Personal Adjustment. New York: Social Science Research Council.

Guttman, L. (1947) The Cornell technique for scale and intensity analysis. Educ. Psychol. Msmt., 7, 247-280.

Jackson, J. E. and Fleckstein, Mary (1957) An evaluation of some statistical techniques used in the analysis of paired comparisons data. Biometrics, 13, 51-64.

Jones, L. V. (1957) Methodology of preference measurement. Quartermaster Food and Container Institute Report P1101 #4. Chicago: Univ. of Chicago.

Kelley, H. H., Hovland, C. I., Schwartz, M. and Abelson, R. P. (1955) The influence of judges' attitudes in three methods of attitude scaling. J. Soc. Psychol., 42, 147-158

Kendall, M. G. (1955) Further contributions to the theory of paired comparisons. Biometrics, 11, 43-62.

Luce, R. D. (1959) Individual choice behavior: A theoretical analysis. New York: Wiley (In press).

Messick, S. (1956) An empirical evaluation of multidimensional successive intervals. Psychometrika, 21, 367-375.

Messick, S. and Abelson, R. P. (1957) Scaling and measurement theory. Rev. Educ. Res., 27, 487-497.

Morris, C. and Jones, L. (1955) Value scales and dimensions. J. abnorm. soc. Psychol., 51, 523-535.

Mosier, C. I. (1940) Psychophysics and mental test theory. Fundamental postulates and elementary theorems. Psychol. Rev., 47, 355-366.

Mosteller, F. (1951) Remarks on the method of paired comparisons: III. A test of significance for paired comparisons when equal standard deviations and equal correlations are assumed. Psychometrika, 16, 207-218.

Mosteller, F. (1958) The mystery of the missing corpus. Psychometrika, 23, 279-289.

Rimoldi, H.J.A., and Hormaeche, Marceva. (1955) The law of comparative judgment in the successive intervals and graphic rating scale methods. Psychometrika, 20, 307-318.

Ross, R. T. A linear relationship between paired comparisons and rank order. (1955) J. exp. Psychol, 50, 352-354.

Scheffé, H. (1952) An analysis of variance for paired comparisons. J. Amer. Stat. Assn., 47, 381-400.

Schwartz, N. and Pratt, Cornelia H. (1956) Simultaneous vs. successive presentation in paired comparison situation. Food Res., 21, 103-108.

Shepard, R. N. (1957) Stimulus and response generalization: a stochastic model relating generalization to distance in psychological space. Psychometrika, 22, 325-345.

Shepard, R. N. (1958) Stimulus and response generalization: Tests of a model relating generalization to distance in psychological space. J. exp. Psychol, 55, 509-523.

Siegel, S. (1956) A method of obtaining an ordered metric scale, Psychometrika, 21, 207-216.

Stevens, S. S. (1957) On the psychophysical law. Psychol. Rev., 64, 153-181.

Stevens, S. S. and Galanter, E. H. (1957) Ratio scales and category scales for a dozen perceptual continua. J. exp. Psychol, 54, 377-411.

Thurstone, L. L. (1927) A law of comparative judgment. Psychol. Rev., 34, 273-286.

Torgerson, W. S. (1952) Multidimensional scaling I. theory and method. Psychometrika, 17, 401-419.

Torgerson, W. S. (1958) Theory and methods of scaling. New York: Wiley.