Colman, A. M., Norris, C. E., & Preston, C. C. (1997). Comparing rating scales of different lengths: Equivalence of scores from 5-point and 7-point scales. *Psychological Reports, 80,* 355-362.

Comparing Rating Scales of Different Lengths: Equivalence of Scores From 5-Point and 7-Point Scales[1, 2]

Andrew M. Colman
Department of Psychology
University of Leicester, UK

Claire E. Norris
Department of Human Communication
De Montfort University, UK

Carolyn C. Preston
Department of General Practice and Primary Health Care
University of Leicester, UK

[1]Preparation of this article was supported by research grant M88 from BEM Research to Andrew M. Colman and Claire E. Norris. We are grateful to Dr Chris Nicklin for his technical advice and assistance, and to Gareth G. Jones of BEM for drawing our attention to the problem discussed in this article.

[2]Address requests for reprints to Andrew M. Colman, Department of Psychology, University of Leicester, Leicester LE1 7RH, UK, or e-mail via the Internet (amc@leicester.ac.uk).

Summary.–Using a self-administered questionnaire, 227 respondents rated service elements associated with a restaurant, retail store, or public transport company on several 5-point and 7-point rating scales. Least-squares regression showed that linear equations for estimating 7-point from 5-point and 5-point from 7-point ratings explained over 85% of the variance and fitted the data almost as well as higher-order polynomials and power functions. In a cross-validation on a new data set, the proportion of variance explained fell to about 76%. Functionally inverse versions of the derived linear equations were calculated for the convenience of researchers and psychometricians.

Circumstances sometimes occur in which researchers or applied psychologists have to compare scores derived from rating scales with different numbers of response categories. In longitudinal research designs in psychology, education, marketing, and many areas of social research, for example, a 5-point scale that has been used for some time may be replaced by a new 7-point scale, or vice versa, and researchers may wish to establish a basis for continuity in order to enable comparisons to be made between the old and the new data. In other circumstances, researchers may wish to compare newly collected 5-point data with 7-point data already published in a journal article or report, or to compare different published sets of 5-point and 7-point data.

A number of studies (reviewed by Cox, 1980) have been conducted to examine the effects of different numbers of response categories on the reliability and validity of rating scales and the response patterns generated by them (e.g., Cicchetti, Showalter, & Tyrer, 1985; Matell & Jacoby, 1971; Schutz & Rucker, 1975). In contemporary psychometric practice, the majority of rating scales, Likert scales, and other attitude and opinion measures contain either five or seven response categories (Bearden, Netmeyer, & Mobley, 1993; Shaw & Wright, 1967). Symonds (1924) was the first to suggest that reliability is optimized with seven response categories, and other early investigations tended to agree (see Ghiselli, 1955, for a comprehensive review of early research). In an influential review article, Miller (1956)

argued that the human mind has a span of absolute judgment that can distinguish about seven distinct categories, a span of immediate memory for about seven items, and a span of attention that can encompass about six objects at a time, which suggested that any increase in number of response categories beyond six or seven might be futile. Odd numbers of response categories have generally been preferred to even numbers because they allow the middle category to be interpreted as a neutral point, and more recent research (e.g. Green & Rao, 1970; Neumann & Neumann, 1981) has tended to reinforce the general preference for 5-point or 7-point scales.

We shall confine our attention to the comparability, equivalence, and estimation–in both directions–between 5-point and 7-point scales, although the mathematical and empirical methods may be generalized to rating scales that differ arbitrarily in numbers of response categories. We shall first discuss naive mathematical solutions, and we shall explain why these solutions are fundamentally untrustworthy. We shall then outline empirical solutions, based on the ratings given by respondents in a large-scale survey of attitudes towards services. Finally, we shall present some recommendations to researchers and practitioners who find themselves confronted with these problems.

Naive Mathematical Solutions

The easiest and most obvious method of estimation, and consequently the one that is probably most widely used, is a simple proportional transformation. This approach involves multiplying each 5-point score by the proportion 7/5 to scale it up to an equivalent 7-point score, or multiplying each 7-point score by 5/7 to scale it down to an equivalent 5-point score. This method of solution can be visualized by imagining an elastic ruler with five equidistant numerals is stretched evenly to fit alongside a longer ruler with seven numerals, or one with seven numerals compressed to fit alongside a ruler with five.

An analogous mathematical solution entails transforming the original 5-point or 7-point scores to standard (z) scores) and then treating them as fully equivalent and comparable. Transformation of raw scores to standard scores is achieved through the relation $z = (x - M)/s$, where x is the raw score, M is the mean of the raw scores, and s is the standard deviation of the raw scores. The standard deviation is the square root of the variance s^2, an unbiased estimate of which is $s^2 = \Sigma[(x_i - M)]/(N - 1)$, where N is the number of raw scores x_i and the summation is over i from $i = 1$ to $i = N$. Standard scores are widely used for comparing raw scores from different distributions, because they are dimensionless quantities with mean and standard deviation equal to 0 and 1 respectively, yet they retain the original shape or mathematical form of the raw score distributions from which they are derived. Standardization is an obvious and natural approach that has proved useful for evaluating empirical data (Rosenthal & Rosnow, 1991), designing experiments (Cohen, 1988), and integrating results from many studies (Hedges & Olkin, 1985), but it has certain drawbacks for comparing data from rating scales of unequal lengths. In particular, it can be used for converting scores only when the mean and standard deviations for the scales are known, and this information is not always available in published and unpublished data that researchers may which to convert.

Although elementary mathematical solutions may be popular in practice, they are likely to yield inaccurate equivalences, because they contain hidden assumptions about human information processing. A purely mathematical approach provides no basis for the choice of suitable parameters for the transformation equation; these can be established only through empirical research. The solution via standardization also rests on implicit assumptions about psychological equivalences between scales of different lengths. How people respond to rating scales with unequal numbers of response categories is a quintessentially *psychological* rather

than a mathematical question, and the aim of this study is to derive the best solution by analyzing data from *empirical research.*

Empirical Study

We obtained responses on a variety of 5-point and 7-point rating scales from 227 respondents throughout England and Wales, 77 men and 150 women aged 20 to "over 60" (in the over-60 range, exact ages were not recorded). The respondents were recruited by a form of snowball sampling with the help of students who volunteered to participate as respondents and to recruit additional respondents in return for course credits. The sample thus consisted of undergraduate students and their friends (some of whom were also undergraduate students) and relatives. Through a self-administered questionnaire, the respondents rated a retail store, restaurant, or public transport company with which they had recent experience. These service categories were chosen on the assumption that all respondents would have used a store, restaurant, or public transport in the recent past, and this turned out to be the case.

The respondents first rated overall service quality ("How would you rate the overall quality of the [store, restaurant, or public transport company]") on a 7-point scale, and they then rated the quality of a key service element associated with that service provider on 5-point and 7-point scales. Different service elements were rated for different service categories: helpfulness of staff (store), competence of staff (restaurant), and availability of information (public transport). The rating scales were presented with the two extremes of the five or seven response categories anchored by either bipolar adjective pairs (e.g., in a scale to rate the *helpfulness of staff,* the anchors were *not at all helpful* at one end of the scale and *extremely helpful* at the other) or by comparisons with the level of service expected (e.g., in the scale relating *helpfulness of staff* to expectations, the anchors were *considerably better than expected* and *considerably worse than expected*. On two-thirds of the rating scales the response categories were displayed as a series of numerals from 1 to 5 or 7, and the respondents were asked to circle or tick an appropriate number. On all other scales the response categories were simply five or seven open bracket pairs, and the respondents were asked to place a tick in the appropriate space. Our aim was to include some commonly used presentation formats and a variety of subject matter with a reasonably representative sample of respondents in terms of sex, age, and geographical distribution.

Results and Analysis

In our analysis of the results, we made comparisons between responses to 5-point and 7-point scales that differed only in number of response categories. The correlation between the 5-point and 7-point scales was high ($r = .921$, $p < .001$). The ratings were analyzed by least-squares regression to determine the best fit of linear, quadratic, third-order polynomial, and power function equations, which are the simplest equations that might reasonably be expected to explain the relationship between the 5-point and 7-point ratings. The results are summarized in Table 1. The R^2 values for the simple proportional and *z* transformations are included for comparison.

Table 1. Simple Proportional Transformation and Least-Squares Regression: Coefficients of Determination for Different Methods (Cross-Validation Data in Parentheses), $N = 227$

	Type of Equation Used for Fitting Data					
Dependent Variable	Simple Proportion	*z* Score	Linear	Quadratic	3rd-Order Polynomial	Power Function
7-point ratings	.841 (.770)	.848 (.773)	.848 (.775)	.848 (.775)	.851 (.784)	.848 (.774)
5-point ratings	.824 (.775)	.848 (.773)	.848 (.769)	.848 (.771)	.849 (.775)	.846 (.768)

The following equations represent the least-squares best fitting linear, quadratic, third-order polynomial, and power functions. In these equations, x represents the observed 5-point or 7-point ratings, y_7 and y_5 represent the estimated 7-point and 5-point ratings respectively, and the numerical estimates of the constants a, b, etc. are derived from the regression analysis and are shown together with the limits of their standard errors of estimate (the probability that an estimated score will fall within one standard error of its predicted value is approximately 68%).

Linear Equations ($y = ax + b$):

$$y_7 = (1.35 \pm .04)x + (.01 \pm .13), \quad [1]$$
$$y_5 = (.63 \pm .18)x + (.50 \pm .08). \quad [2]$$

The coefficient of determination for Equation 1 is $R^2 = .848$, and for Equation 2 it is also $R^2 = .848$; in each case 84.8% of the variance in ratings is accounted for by the linear equation.

Quadratic Equations ($y = ax^2 + bx + c$):

$$y_7 = (-.02 \pm .03)x^2 + (1.44 \pm .21)x - (.13 \pm .29), \quad [3]$$
$$y_5 = (-.00 \pm .01)x^2 + (.66 \pm .09)x + (.45 \pm .17). \quad [4]$$

The coefficient of determination for Equation 3 is $R^2 = .848$, and for Equation 4 it is $R^2 = .848$, indicating that the least-squares fit is no better than for the linear equations.

Third-Order Polynomial ($y = ax^3 + bx^2 + cx + d$):

$$y_7 = (-.06 \pm .03)x^3 + (.53 \pm .26)x^2 - (.04 \pm .74)x + (1.00 \pm .61), \quad [5]$$
$$y_5 = (-.01 \pm .01)x^3 - (.11 \pm .08)x^2 + (.26 \pm .30)x + (.83 \pm .32). \quad [6]$$

The coefficient of determination for Equation 5 is $R^2 = .851$ and for Equation 6 it is $R^2 = .849$. The slight increase over the coefficients for Equations 1 to 4 is inconsequential: higher-order polynomials necessarily provide better least-squares fits to virtually all data sets than lower-order polynomials, because they contain more terms and parameters.

Power Function ($y = ax^b$):

$$y_7 = (1.34 \pm .07)x^{(1.00 \pm .04)}, \quad [7]$$

$$y_5 = (.99 \pm .05)x^{(.82 \pm .30)}. \quad [8]$$

The coefficient of determination for Equation 7 is $R^2 = .848$, and for Equation 8 it is $R^2 = .846$. These figures show that the power function equation accounts for about 85% of the variance in estimating 5-point ratings from 7-point ratings, and vice versa.

The regression equations (1) to (8) together with the simple proportional and *z* score transformations were cross-validated for goodness of fit with a new data set. These data were from 224 of the participants in the original study responding to questions about a different service element (promptness of service). The R^2 values for the goodness of fit to this new set of cross-validation data are presented in parentheses in Table 1. As expected with cross-validation, the values of R^2 are lower than for the original set of data (regression equations almost invariably fit the data from which they are derived better than cross-validation data). Also, the correlation between the 5-point and 7-point scales is slightly lower for these data ($r = .879$, $p < .001$). However, the pattern of results is similar to the original data set, with all equations fitting reasonably well (accounting for 76.8% – 78.4% of the variance). The simple proportional transformation again fit more poorly than the other equations.

Inverse Linear Equations

The linear, quadratic, third-order polynomial, and power function equations generated estimates that did not differ meaningfully from one another in accuracy: the lowest coefficient of determination for the original set of data was $R^2 = .846$ and the highest was $R^2 = .851$. In the light of these findings, the most suitable method of estimation is probably best chosen with the help of Occam's razor. The simplest is the linear transformation, and it seems the most sensible choice for general use.

However, for practical applications it is desirable to have a pair of equations with an inverse functional relationship to each other, that is, an equation for estimating 7-point from 5-point ratings that is an inverse function of the equation for estimating 5-point from 7-point ratings. Equations 1 and 2 do not have this inverse relation to each other. For instance, if a 7-point rating is estimated using Equation 1 from a 5-point rating *x* and the resulting estimation is then inserted into Equation 2 (to estimate its 5-point equivalent), the result will not be identical to the value of the original 5-point rating *x*.

A pair of inverse linear equations can be calculated by averaging the derived regression equations. Rewriting Equations 1 and 2 uniformly, using x_7 to represent 7-point ratings and x_5 to represent 5-point ratings,

$$x_7 = (1.35 \pm .04)x_5 + (.01 \pm .13), \quad [9]$$
$$x_5 = (.63 \pm .18)x_7 + (.50 \pm .08). \quad [10]$$

From Equation 10,

$$(.63 \pm .18)x_7 = x_5 - (.50 \pm .08). \quad [11]$$

With due attention to the error terms, this leads to

$$x_7 = (1.59 \pm .45)x_5 - (.79 \pm .26). \quad [12]$$

Taking Equations 9 and 12, averaging, and inverting to provide two equations, then,

$$x_7 = (1.47 \pm .23)x_5 - (.40 \pm .15), \quad [13]$$
$$x_5 = (.68 \pm .10)x_7 + (.27 \pm .11). \quad [14]$$

Equations 13 and 14 provide a fully invertible method of estimating 7-point from 5-point and 5-point from 7-point ratings with negligible loss of accuracy compared to the least-squares regression equations. In fact, the R^2 values for the revised equations are both .843 for the original data, showing that they account for 84.3% of the variance. For the cross-validation data, the values are .764 for the conversion from a 7-point to a 5-point scale, and .754 for the conversion from a 5-point to a 7-point scale.

Discussion and Conclusions

Previous research does not appear to have focused on the problems of comparison, equivalence, and estimation of scores derived from rating scales with unequal numbers of response categories or alternatives. Such problems are ubiquitous in a wide variety of pure and applied research, and non-empirical solutions are inadequate. They are analogous to psychophysical problems, requiring solutions based on empirical information about how people respond to rating scales that differ only in their numbers of response categories.

The results showed that linear regression equations gave results virtually equivalent to those derived from more complicated transformations. In hindsight this is not surprising. Psychophysical relations between the magnitude of sensations and the physical intensity of their corresponding stimuli have usually been found to be best described by logarithmic or power functions (Stevens, 1975). In the case of rating scales with unequal numbers of response categories, the relationship between the two variables is a psychological rather than a psychophysical relation, and some simple relationship could perhaps have been anticipated.

The multiplicative constant a in the linear equation $y = ax + b$ turned out to be close to 7/5 in Equation 13 and to 5/7 in Equation 14. However, the linear regression equations are preferable to the simple proportional transformation (multiplying by 7/5 or 5/7), because they are empirically derived, include the extra specification of an additive constant, and provide error terms. Straightforward z transformations fit the data as well as the linear transformations, but they can be applied only when data are available from which to estimate the variance or standard deviation of the untransformed scores, and such data are not always provided in summaries of data collected in the past. When standard deviations are unavailable, z scores cannot be calculated, whereas transformations via the inverse linear equations derived in this article may still be used for converting scores.

Although 5-point and 7-point rating scales are by far the most common, other scale lengths are sometimes used. Further research is required to determine whether the conclusions reported in this article apply more generally to other scale lengths. Meanwhile, the inverse Equations 13 and 14 for the comparison of 5-point and 7-point data are recommended for the estimation of equivalences.

References

Bearden, W. O., Netmeyer, R. G., & Mobley, M. F. (1993) *Handbook of marketing scales: multi-item measures for marketing and consumer behavior research.* Newbury Park, CA: Sage.

Cicchetti, D. V., Showalter, D., & Tyrer, P. J. (1985) The effect of number of rating scale categories on levels of inter-rater reliability: a Monte-Carlo investigation. *Applied Psychological Measurement, 9,* 31-36.

Cohen, J. (1988) *Statistical power analysis for the behavioral sciences* (2nd ed.). Hillsdale, NJ: Erlbaum.

Cox, E. P. (1980) The optimal number of response alternatives for a scale: a review. *Journal of Marketing Research, 17,* 407-422.

Ghiselli, E. E. (1955) *The measurement of occupational aptitude.* Berkeley, CA: Univers. of California.
Green, P. E., & Rao, V. R. (1970) Rating scales and informational recovery – how many scales and response categories to use? *Journal of Marketing, 34,* 33-39.
Hedges, L. V., & Olkin, I. (1985) *Statistical methods for meta-analysis.* Orlando, FL: Academic Press.
Matell, M. S., & Jacoby, J. (1971) Is there an optimal number of alternatives for Likert scale items? Study 1: reliability and validity. *Educational and Psychological Measurement, 31,* 657-674.
Miller, G. A. (1956) The magical number seven, plus or minus two: some limits on our capacity for processing information. *Psychological Review, 63,* 81-97.
Neumann, L., & Neumann, Y. (1981) Comparison of six lengths of rating scales: students' attitudes toward instruction. *Psychological Reports, 48,* 399-404.
Rosenthal, R., & Rosnow, R. L. (1991) *Essentials of behavioral research: methods and data analysis* (2nd ed.). New York, NY: McGraw-Hill.
Schutz, H. G., & Rucker, M. H. (1975) A comparison of variable configurations across scale lengths: an empirical study. *Educational and Psychological Measurement, 35,* 319-324.
Shaw, M. E., & Wright, J. M. (1967) *Scales for the measurement of attitudes.* New York, NY: McGraw-Hill.
Stevens, S. S. (1975) *Psychophysics.* New York, NY: Wiley.
Symonds, P. M. (1924) On the loss of reliability in ratings due to coarseness of the scale. *Journal of Experimental Psychology, 7,* 456-461.