Perception & Psychophysics
1982, 32 (3), 293-296

On rescaling data to fit the model and concluding that the model fits: A note on monotone transformation

MICHAEL H. BIRNBAUM
University of Illinois, Urbana-Champaign
Champaign, Illinois

Carterette and Anderson (1979) advocate a procedure for monotone transformation of data obtained with factorial experiments to determine whether deviations from theoretical predictions can be removed by weak monotonic transformation of the data. To minimize theory-data discrepancies, their procedure employs a separate transformation for each replicate. They propose to perform ANOVAs on the transformed scores as though no parameters were estimated in the transformation procedure, and to draw inferences concerning the psychological process from the value of the calculated F for interactions.

This note questions their approach and discusses alternative procedures for transformation and for testing theories on the basis of ordinal data. It argues that error theory and statistical theory are separate issues, and should be treated as such.

A Model of Transformation

Carterette and Anderson (1979) do not state their theory explicitly, and therefore certain aspects of their paper are unclear. However, they appear to entertain the following model:

$$T_k(R_{ijk}) = a_i + b_j + e_{ijk}, \quad (1)$$

where R_{ijk} is the response in the i^{th} row and j^{th} column of the factorial design for the replicate k, T_k is a set of weak monotone functions, estimated separately for each replicate, a_i and b_j are estimated parameters that are assumed to be independent of the replicate, and e_{ijk} is an error term. Unfortunately, Carterette and Anderson make few statements about the error term, but they use standard analysis of variance on the transformed scores, so it appears that they accept the usual assumptions of normality and of homogeneity of variance and covariance among the errors.[1] They present no information concerning tests of any of these assumptions for their data.

The theories considered by Carterette and Anderson (1979) imply no interaction between factors. The transformations, T_k, were designed to reduce the proportion of variance in the interactions relative to main effects. After transforming the data to reduce an index of stress, they calculated the $\hat{F}$ ratio for the interaction in the transformed scores. In one case, the $\hat{F}$ ratio after transformation was greater than that before transformation. Had the $\hat{F}$ itself been used as the criterion to minimize, it would have grown smaller after transformation.

Carterette and Anderson (1979) present no evidence that the procedure they advocate leads to a calculated $\hat{F}$ (after transformation) that obeys the F distribution with the stated degrees of freedom. Before concluding that the procedure of Carterette and Anderson is "valid," as they claim, it should be demonstrated by derivation or Monte Carlo methods that the calculated $\hat{F}$ under the null hypothesis does, indeed, have the appropriate F distribution.

Without evidence that the procedure leads to a test statistic that has the assumed distribution or that the transformed scores obey the ANOVA assumptions, a skeptic should reserve judgment concerning inferences drawn from the procedure of Carterette and Anderson (1979).

Deeper Issues

A deeper issue is the extent to which the replication factor provides constraint to prevent possibly inappropriate transformation. In other words, is there any information in the response variability above and beyond the means or medians that can prevent inappropriate transformation? Four responses to this question follow:

First, any function that is ordinally additive will still be rescalable to additivity. For example, if the "true" model were

$$\psi_{ijk} = a_i b_j e_k,$$

where $R_{ijk} = T^*(\psi_{ijk})$ and T^* is a monotonic function, it should be clear that the transformation of Equation 1 would rescale to additivity. The values of a_i and b_j would depend on the model assumed.

Second, the procedure advocated by Carterette and Anderson (1979) would succeed in removing the interaction in Table 1, which shows a clear violation of independence. The right half of Table 1 shows that the transformed scores are perfectly additive. With each replicate resembling Table 1, it should be clear that averaging transformed data over replications is no prophylactic for degenerate and therefore inappropriate transformation. Fortunately, Carterette and Anderson advocate plotting the original data as well as the transformed data, and the deviation in Table 1 would be noticed in plots of the original

Thanks are due Jerome Busemeyer and Barbara Mellers for their comments on an earlier draft. The author's mailing address is: Department of Psychology, University of Illinois, 603 East Daniel, Champaign, Illinois 61820.

0031-5117/82/090293-04$00.65/0

Table 1
Example of Degenerate Transformation

	Hypothetical Data			Rescaled Data	
	b_1	b_2		b_1	b_2
a_1	2	4	a_1	2	4
a_2	5	8	a_2	6	8
a_3	6	7	a_3	6	8
a_4	9	10	a_4	10	12

Note–The rescaled data are a weak monotone function of the hypothetical data. If this pattern occurs in each replicate, the fit to the additive model will be perfect, despite ordinal violations of additivity in every replicate.

data. However, the use of replications would not prevent such degeneracy. Indeed, use of a separate transformation for each replicate could, in fact, encourage degenerate solutions, in which two scale values close in value would be squashed into the same value to improve the overall fit. Busemeyer (1980) has noted this problem and illustrated it with another example.

Third, there exists a simple argument to show that transformation of the raw data and of averaged data should lead to the same conclusions, if, at least, the median is accepted as the average. This seems a mild requirement, since means and medians are usually monotonically related for judgment tasks. If T is a monotonic transformation, and M refers to the median, then $T[M(x)] = M[T(x)]$; that is, the median of the transformed scores will be the same as the transformation of the median. Therefore, whether the scores are transformed separately and then averaged or averaged and then transformed should make no difference. However, when *different* monotonic transformations are permitted for each replicate, then the transformed scores are *not* necessarily a monotonic function of the original data. Under these conditions, the mean transformed scores have a theoretical status different, relative to the original data, from that when a single transformation is applied. Therefore, if a transformation, T, exists that renders the medians parallel, then a single transformation of the raw data that renders the medians of the transformed data equally parallel also exists. Therefore, if a single transformation successfully removes the interaction in the medians, it does not seem sound reasoning to argue that Carterette and Anderson's (1979) procedure, which uses *different* transformations for different replicates, gains additional *constraint* that could have prevented inappropriate transformation.

Fourth, additional constraint can be gained from response variability by explicitly transforming data to fit a distributional theory, as outlined in the next section.

A DISTRIBUTIONAL THEORY AND TRANSFORMATION TECHNIQUE

The following procedure, which was briefly outlined by Birnbaum (1979), does gain additional constraint from the data, by assuming that the errors have the same distribution for all cells in the design. This assumption is forced into the analysis as an additional constraint upon the transformation. Of course, the theory may not be correct, but that is a risk with any theory.

The idea is that the law of categorical judgment (Torgerson, 1958), related to Thurstone's laws and signal detection theory, can be extended to judgments obtained in factorial designs as follows: Let $P_{ijm} = \text{Prob}(R_{ij} \geqslant X_m)$, where X_m is a response value. Then a reasonable model can be written

$$P_{ijm} = F[(a_i + b_j - t_m)/\sigma_{ij}], \tag{2}$$

where F is a strictly monotonic distribution function (in Thurstone's law, it would be the cumulative normal), a_i and b_j are as before, t_m is the subjective value of a response $= X_m$, and σ_{ij} is the psychological dispersion for this cell in the design. If it is assumed that σ_{ij} is a constant for all cells, then Equation 2 becomes additive in three factors: rows, columns, and response value. Monotone scaling programs can be used to solve for a_i, b_j, t_m, and F^{-1}. The relationship between subjective value and response can be represented as the function $X_m = J(t_m)$.

A more general model can be written as

$$P_{ijm} = F[\psi_{ij} - t_m/\sigma_{ij}], \tag{3}$$

where ψ_{ij}, which represents the subjective impression, is not constrained to be additive. A special case of Equation 3 would assume $\sigma_{ij} = 1$, as follows:

$$P_{ijm} = F[\psi_{ij} - t_m]. \tag{4}$$

This model assumes that the distributions on the subjective continuum are identical for all of the stimulus combinations. One could derive the values of ψ from this model and plot them to check if the assumption of homogeneity leads to parallelism. If it did not, one could question either additivity or homogeneity.

Equations 2, 3, and 4 are models that incorporate the distribution of errors into the judgment theory. Each of the equations leads to a monotone rescaling procedure. However, these models are best regarded as *theories*, rather than *methods* for statistically processing the data.[2] It should be clear that there will be an indeterminacy of error model and judgment model. Thus, if the assumption of homogeneous error

variance is consistent with additivity, then the assumption that error variance is proportional to ψ would lead to a multiplying model. This indeterminacy is analogous to the case V vs. case VI dispute in Thurstone scaling (Bock & Jones, 1968).

Example

Tables 2 and 3 illustrate the application of Equation 4 with hypothetical data. The medians of the raw data show a bilinear interaction between row and column (Table 3) that would be consistent with a multiplying model but not an additive model. However, when the data are rescaled to homogeneity (Table 2 is rescaled to parallelism), then the values of ψ derived from Equation 4 are additive (the right-hand portion of Table 3). This example used the cumulative normal function for F, but the procedure assumes only that F is a monotone function, and it allows estimation of the function. This example illustrates the indeterminacy between choice of model and the error theory.

SCALE-FREE TESTS

A far better procedure for assessing models, such as the parallel-averaging model for impression formation or the size-weight illusion, is by means of scale-free tests, as in Birnbaum (1974, 1982) and Birnbaum and Veit (1974b). These studies provide data that would yield the same conclusion regarding the additive model, irrespective of monotone transformation of the data. Birnbaum (1974) showed that the parallel-averaging model of impression formation can be rejected. Similarly, Birnbaum and Veit (1974b) found that the size-weight illusion shows systematic violations of the parallelism-averaging model.

The scale convergence criterion (Birnbaum, 1974, 1982) also constrains monotone transformation (Birnbaum & Veit, 1974a). Scale convergence also provides evidence against interpretations derived from scale-dependent research reviewed by Anderson (1979) (see Birnbaum, 1982). Anderson's (1977, 1979) suggestion that impression formation and the size-weight illusion can be approximated by parallel-averaging models is probably restricted to scale-dependent studies using small designs that cover a narrow range of values.

The studies reviewed by Anderson (1979) are termed "scale-dependent" because the conclusions regarding impression formation and the size-weight illusion are not invariant under monotone transformation of the data. In Birnbaum's (1974, 1982) scale-free approach, violations of the additive or parallel-averaging model cannot be attributed to nonlinearity in the response function.

Table 2
Hypothetical Results of Factorial Experiment

	Response						
SC	12	25	50	100	200	400	800
a_1b_1	.16	.50	.84	.97	.97	.97	.97
a_1b_2	.03	.16	.50	.84	.97	.97	.97
a_1b_3	.03	.03	.16	.50	.84	.97	.97
a_2b_1	.03	.16	.50	.84	.97	.97	.97
a_2b_2	.03	.03	.16	.50	.84	.97	.97
a_2b_3	.03	.03	.03	.16	.50	.84	.97
a_3b_1	.03	.03	.16	.50	.84	.97	.97
a_3b_2	.03	.03	.03	.16	.50	.84	.97
a_3b_3	.03	.03	.03	.03	.16	.50	.84

Note–SC = stimulus combination. Each entry is the proportion of responses to the row stimulus combination less than the column response (proportions more extreme than .03 and .97 have been tied to these values). Equation 4 is revised: $P_{ijm} = F(t_m - \Psi_{ij})$.

Table 3
Raw and Transformed Medians for Hypothetical Data

	Column					
	Raw Data			Transformed		
Row	b_1	b_2	b_3	b_1	b_2	b_3
a_1	25	50	100	−2	−1	0
a_2	50	100	200	−1	0	1
a_3	100	200	400	0	1	2

Note–For Equation 4, values of t_k *are −3, −2, −1, 0, 1, 2, and 3 for 12, 25, 50, 100, 200, 400, and 800, respectively. Thus, this hypothetical example assumes that J is an exponential function and that errors are homogeneously distributed.*

CONCLUDING COMMENTS

The transformation procedure of Carterette and Anderson (1979) is an attempt to develop a statistical test for nonmetric scaling. However, the procedure seems dubious because its statistical properties have not been demonstrated. Furthermore, the procedure does not truly use the distribution of responses as a constraint for transformation. Moreover, the focus on statistical tests of fit may be misguided. Given any true departure from a model, no matter how small, it should be possible to collect enough data to reject the model. Scientific decisions about the success or failure of a transformation are (and should be) based on examination of the critical properties of the data, for example, examination of the graphical properties of the theory and data, rather than any overall index of fit or deviations.

The models and transformation procedure described under Equations 2, 3, and 4 are theories that attempt to specify the distribution of errors on the subjective continuum. In testing these models, it should be possible by means of graphical tests to discover whether homogeneity of error variance, for example, is empirically compatible with additivity of

effects. As a by-product of the rescaling, the distribution function F can be estimated, and it should be possible to assess the fit of particular functions such as the cumulative normal distribution. However, few judgment theories offer implications for the error distribution; therefore, there is an inherent indeterminacy of algebraic model and error theory.

To resolve indeterminacies of testing models when monotone transformation is permitted, it therefore proves necessary to provide additional leverage, such as the constraint of scale convergence or the scale-free tests.

REFERENCES

ANDERSON, N. H. Note on functional measurement and data analysis. *Perception & Psychophysics*, 1977, **21**, 201-215.

ANDERSON, N. H. Algebraic rules and psychological measurement. *American Scientist*, 1979, **67**, 555-563.

BIRNBAUM, M. H. The nonadditivity of personality impressions. *Journal of Experimental Psychology*, 1974, **102**, 543-561. (Monograph)

BIRNBAUM, M. H. Reply to Eisler: On the subtractive theory of stimulus comparison. *Perception & Psychophysics*, 1979, **25**, 150-156.

BIRNBAUM, M. H. Controversies in psychological measurement. In B. Wegener (Ed.), *Social attitudes and psychophysical measurement*. Hillsdale, N.J: Erlbaum, 1982.

BIRNBAUM, M. H., & VEIT, C. T. Scale convergence as a criterion for rescaling: Information integration with difference, ratio, and averaging tasks. *Perception & Psychophysics*, 1974, **15**, 4-15. (a)

BIRNBAUM, M. H., & VEIT, C. T. Scale-free tests of an additive model for the size-weight illusion. *Perception & Psychophysics*, 1974, **16**, 276-282. (b)

BOCK, R. D., & JONES, L. V. *The measurement and prediction of judgment and choice*. San Francisco: Holden-Day, 1968.

BUSEMEYER, J. R. Importance of measurement theory, error theory, and experimental design for testing the significance of interactions. *Psychological Bulletin*, 1980, **88**, 237-244.

CARTERETTE, E. C., & ANDERSON, N. H. Bisection of loudness. *Perception & Psychophysics*, 1979, **26**, 265-280.

TORGERSON, W. S. *Theory and methods of scaling*. New York: Wiley, 1958.

NOTES

1. Some procedures for transformation are based on the model

$$T_k[R_{ijk}] = a_{ik} + b_{jk} + e_{ijk}, \quad (1a)$$

where the terms are defined as in Equation 1, except that different row and column scale values are now estimated for each replicate. This procedure is even more liberal in terms of permitting more parameters to be estimated from the data. Although the error terms are unlikely to be homogeneous in distribution, they seem more likely to be uncorrelated across replicates with this procedure.

2. Two practical problems in the use of Equations 2, 3, and 4 will arise: (1) How many levels of X_m should be chosen? (2) How should one deal with estimated proportions? To some extent, answers to these questions will depend on the richness of the data set. For most cases of hypothetical data studied, about 7-14 values of X_m seem to work well (fewer for smaller data sets). Extreme proportions (perhaps $\leq .03$ and $\geq .97$ could be tied (set equal to .03 and .97), with the monotone program permitted the option of breaking ties. The proportions should be based on enough data so that, given the standard errors of the estimates, the rank order of P_{ijm} in the sample would be unlikely to differ from that in the population.

(Manuscript received May 13, 1981;
revision accepted for publication June 17, 1982.)