Bulletin of the Psychonomic Society
1979, Vol. 14 (4), 307-310

A note on monotonic transformations in the context of functional measurement and analysis of variance

DAVID V. BUDESCU and THOMAS S. WALLSTEN
University of North Carolina at Chapel Hill, Chapel Hill, North Carolina 27514

The functional measurement approach (Anderson, 1977) technically is based on factorial designs, quantitative response measures, and monotonic transformations, with the analysis of variance (ANOVA) as the usual statistical tool. This note discusses problems that often occur when utilizing monotonic data transformations in this context. Illustrative examples are presented, and a procedure is suggested to alleviate the problems. The procedure uses expected normal scores, or inverse normal scores, or random normal scores to simultaneously provide monotonic transformations and suitable use of ANOVA in testing functional measurement models. A consequence of the procedure is that the impact of measurement error on model validation and scale value estimation is reduced.

Functional measurement (FM) has been applied in a variety of experimental situations since it was first proposed by Anderson (1962). These applications have provided an extensive and highly consistent body of knowledge related to the process of information integration (for a review of the empirical findings, see Anderson, 1974).

The uniqueness of FM is in its realization that "the hypothesized integration function could itself be used as the base and frame for scaling subjective values" (Anderson, 1976, p. 677). Technically, this approach relies on factorial designs, quantitative response measures, and (optionally) monotonic rescaling procedures (Anderson, 1970). In practice, the usual analysis of variance (ANOVA) procedures are used to test the fit of the assumed models and to estimate the psychological values of stimuli. It has been claimed that FM has advantages over methods advocating preliminary scale validation (Anderson, 1971, 1977), and that it also offers a test of fit superior to the correlational and scatterplot methods (Anderson, 1977; Birnbaum, 1973). If these statements and the logic of FM are correct, then this approach is probably the only technique of model evaluation based on a well established error theory (borrowed from ANOVA) and on computational techniques well known and easily available.

The purpose of this note is to consider problems that arise when using monotonic transformations in the context of FM and ANOVA. We will consider the objectives of the transformations, the conditions under which the transformations are applied, and their uniqueness features. In FM there are two possible motivations for monotonic transformations: model validation and scale value estimation. Unfortunately, these two goals sometimes conflict, both on theoretical grounds and in terms of data analysis and interpretation. Following a brief discussion of monotonic transformations, we will propose a procedure for dealing with such transformations, if and when they are needed.

If the results of the analysis of the original scores do not support a particular hypothesized model (a pattern of interactions differing from the one predicted under the model is obtained; Anderson, Note 1, p. 78), one or more of the following reasons must be considered: (1) The data as collected do not conform to the basic assumptions of ANOVA, and the probability inferences are at least inexact and possibly misleading. (2) The original measurement scale is not linearly related to the underlying variable of interest, and therefore real effects are masked and artificial effects appear (see Birnbaum, 1974, for a discussion of the distinction among the overt response, the underlying variable, and the relation between the two). (3) The hypothesized model is incorrect.

If the researcher thinks Reasons 1 and/or 2 are the reasons for the failure of the tests, then he or she may look for transformations that will eliminate the problems. Three possible objectives for such transformations are: (1) to maximize conformity of the data to the assumptions of the statistical analysis; (2) to optimize an objective index under the assumed model (e.g., minimization of the sum of squared deviations between observed and predicted values); and (3) to conform to requirements based on prior theoretical or experimental knowledge about the nature of the data in that particular field. We will concentrate here primarily on

This research was supported in part by NSF Grant BNS76-20759. The authors are indebted to Amnon Rapoport and Warren Sarle for many useful comments on earlier drafts of this paper. Address correspondence to either author at Psychometric Laboratory, Davie Hall 013 A, University of North Carolina, Chapel Hill, North Carolina 27514.

0090-5054/79/100307-04$00.65/0

the empirically based transformations prescribed by Criteria 1 and 2 and only briefly discuss Criterion 3.

Variance Stabilizing Transformations

Functional dependencies between the means and the variances of the data, violating assumptions of ANOVA, are often encountered in psychological research (e.g., with magnitude estimates or reaction times). If the nature of the mean-variance relationship can be characterized, then a unique transformation that will stabilize the variances can be found (Bartlett, 1947; Smith, 1976). In many cases these transformations also improve the normality of the data (Curtis, 1943). In addition to the statistical implications, a particular mean-variance relation often implies a specific distribution underlying the process (e.g., Poisson, binomial, etc.), and this may be a source for theoretical insight (Smith, 1976).

Transformations to Achieve Response Scale Linearity

It is more complicated to find a transformation that will result in a linear scale, as needed in order to apply the FM technique (see Anderson's "parallelism theorem," 1977, p. 202). First, it is difficult to verify that a scale is indeed linear (Torgerson, 1954, p. 82). Second, when a "scale stabilizing" transformation, determined by the nature of the model, is applied and the hypothesized integration function is supported, one still cannot be sure that the scale is linear. It is possible that the transformation introduces a different kind of bias that artificially supports the model (analogous to the case in which the untransformed data artificially cause rejection of the model). Next, although a transformation that improves the fit of the data to the model can always be found, the index of fit will not necessarily exceed some predetermined value or significance level. Finally, and often of substantial importance, there may exist more than one transformation that will improve the fit equally well.

Frequently, monotonic transformations are sought that maximize fit to a particular model, with the assumption that such a procedure entails a more nearly linear scale. A major problem with the "maximal fit" transformations is that they can lead to self-contradictory or indeterminable situations, as in the following two examples. Consider, first, the intriguing case in which each of two competing models (e.g., a distributive and a dual distributive) have enough prior support in the empirical and theoretical literature to justify a "critical test." A factorial design is obtained by systematically manipulating three factors, A, B, and C, and the following pattern of significance tests is obtained: The effects A, B, C, AB, BC, and ABC are all significant, whereas AC is not significant. This pattern of results supports neither of the models under consideration. However, it may be possible to find two transformations, f_1 and f_2, one maximizing the fit to, and leading to validation of, the distributive model, $f_1(R) = \psi_1(A)\psi_2(B) + \psi_2(B)\psi_3(C)$, and the other maximizing the fit to, and leading to validation of, the dual distributive mode, $f_2(R) = \psi_4(A)\psi_5(B) + \psi_6(C)$, thus leaving the matter as indeterminate as before. Moreover, researchers holding different ideas will adopt either f_1 or f_2 as the "correct transformation."

Relations between the Two Classes of Transformation

The two classes of transformation, variance stabilizing and scale stabilizing, frequently prescribe different transformations, since the optimal policies under these objectives are not equivalent, as has already been noted by Bartlett (1947, pp. 41, 50) and by Rapoport and Wallsten (1972, p. 147). This is a disturbing situation. The second example will illustrate the possible contradictions.

Consider a 4 by 4 factorial orthogonal design in which the simple additive model is to be tested. Random samples of size 18 per cell were obtained from each of four different normal populations, and the model was tested for each population. The populations were identical in terms of the cell means, which are shown in Table 1, but differed in terms of the variance. The four mean-variance relations are specified in the following four equations: $\sigma_{ij}^2 = \mu_{ij}$, $\sigma_{ij}^2 = (.25\mu_{ij})^2$, $\sigma_{ij}^2 = (.30\mu_{ij})^2$, and $\sigma_{ij}^2 = (2\mu_{ij})/(n-1)$. All of the observations in each of the four cases were subjected to each of the following transformations: $x'_{ij} = \sqrt{x_{ij}}$, $x'_{ij} = \ln(x_{ij})$, $x'_{ij} = \log(x_{ij})$, $x'_{ij} = \ln(x_{ij} + 1)$, $x'_{ij} = \log(x_{ij} + 1)$, and $x'_{ij} = 1/x_{ij}$.

The original and transformed scores were each analyzed by a regular ANOVA, with the relevant results summarized in Table 2. The entries in the table are the probabilities of obtaining the observed interactions under the hypothesis of no true interaction. First note that on the original scales, the interaction was always significant (or very close to significant) at the traditional .05 level for each of the five sets of data. The italicized entries in each column are the probabilities of chance interaction obtained after application of the correct "variance stabilizing" transformation. Note that in each column there is at least one transformation that improves the additivity of the data and thus fits the model at least as well as the "correct" transformation. The point is clearly demonstrated: An optimal variance stabilizing transformation (Bartlett, 1947; Smith, 1976) is not necessarily optimal for fitting the hypothesized model.

Table 1
True Population Means on the Measured Scale

Levels of Factor B	Levels of Factor A			
	1	2	3	4
1	10	12	14	16
2	12	14	16	19
3	13	15	17	20
4	15	18	21	25

Table 2
Probabilities of Interactions Under the Null Hypothesis of No Interaction in the ANOVA Tests of the Original and Transformed Scores

	σ^2_{ij}			
Scale	μ	$(.25\mu)^2$	$(.3\mu)^2$	$(2\mu^2)/(n-1)$
x	.015	.014	.027	.066
$\sqrt{x}$	*.106*	.057	.068	.326
ln(x)	.348	*.152*	*.140*	*.999*
log(x)	.343	*.154*	*.145*	.999
ln(x + 1)	.289	*.141*	*.125*	.999
log(x + 1)	.287	*.135*	*.124*	.999
1/x	.420	.999	.382	.075

Note–Differences between significance levels of ln(x) and log(x), and ln(x + 1) and log (x + 1) are due to rounding errors. Italicized entries are the probabilities of chance interaction obtained after application of the correct "variance stabilizing" transformation.

AN ALTERNATIVE APPROACH

It is evident from the discussion and examples above that additional criteria and a better defined procedure are required for choosing among transformations. In other words, a decision rule for choosing among transformations satisfying different criteria is needed. Such a decision rule should satisfy the following objectives: (1) The procedure should enable one both to test the model and to estimate its parameters. (2) The transformation should be independent of the particular model being tested. (3) It should not capitalize on chance.

The following procedure satisfies these objectives. (1) Inspect the data for conformity to ANOVA assumptions. Assuming a sufficiently large sample, this is best done by plotting the variances as a function of the means, because there is no generally accepted test of normality, and most tests of homogeneity of variance are highly sensitive to nonnormality or to differences of the means. (2) If a dependency between the means and variances is revealed, transform the data by the recommended transformation (Bartlett, 1947; Smith, 1976), as discussed above. Generally, ANOVA is described as a highly robust technique, but it is not universally so (see Glass, Peckham, & Sanders, 1972, for a detailed discussion), and the variance stabilizing transformations are suggested in cases of extreme and clear violations of assumptions.

(3) Analyze the data by ANOVA, and if the model is not rejected, go to Step 6 below.

(4) If at this point the model is rejected (see conditions in Anderson, Note 1), transform the data to ranks, proceed with either of the two following classes of transformations, and analyze the transformed data by ANOVA. (a) Transform the ranks to expected normal scores or to inverse normal scores (Bradley, 1968, Chapter 6). In the first case, replace the rth smallest observation in the pooled sample of N observations by the expected value of the rth smallest observation in a sample of size N from a standard normal population. In the second case, replace the rth smallest observation in the pooled sample of N observations with the standard normal deviate whose cumulative probability is r/(N + 1). Note that the new transformed scores are not random variables, but rather are fixed scores, and this fact introduces a certain bias in the tests. However, exact statistics exist for the ANOVA of expected normal scores and of inverse normal scores, based on the randomization principle (see Bradley, 1968, Chapter 6), and they can be applied for small samples. For large enough samples, the two tests become identical, and the F distribution can be used as an asymptotic approximation to them. (b) Transform the ranks to random normal scores (Bradley, 1968, Chapter 6). Replace the rth smallest observation in a pooled sample of N observations by the rth smallest observation in a random sample of N observations from a standard normal distribution. In this case the new transformed scores are random variables that conform perfectly with the assumptions of ANOVA, but they contain an additional source of chance, unrelated to the original data.

Either one of the two transformations allows for exact tests for main effects, simple effects, interactions, or bilinear interactions (after alignment), and so on, and can be interpreted in the same manner as tests performed on the original scores (for a review of these techniques, see Marascuilo & McSweeney, 1977). For large samples regular ANOVA tests can be performed, because the transformed scores are asymptotically normal. If at this point the model is not rejected, proceed to Step 6.

(5) If at this point the model is rejected, and if it is a two- or three-variable polynomial, test the necessary axioms of the model according to conjoint measurement theory (CMT) (Krantz & Tversky, 1971). No general error theory currently exists for testing these axioms in a probabilistic way, but the usual nonparametric indices used to test them (Holt & Wallsten, Note 2) can usually be interpreted. Note that the CMT axioms are used at this point not in order to test the model (it was already rejected), but, rather, in order to reveal the exact reasons for its failure.

(6) If the model has been supported, transform the original data to maximize their fit, if necessary, and estimate scale values under the already validated model. Note that now we transform the data according to a criterion that was judged inappropriate for model validation. Initially, we did not want to validate the model on the basis of a model-dependent transformation, but once the model is validated, such a transformation becomes acceptable (although not necessary). This distinction between validation and estimation procedures is parallel to the distinction between qualitative and quantitative properties of a model (Krantz & Tversky, 1971).

To summarize, the proposed procedure provides for a rank transformation, although a second transformation is often useful because of data-analytic considerations. Note that the composite of the two transformations does not depend on the particular scale of measurement, and the scale values have well specified uniqueness properties.

The preceding procedure is based on the following principles of FM, ANOVA, and CMT: (1) Numerical data collected in a well designed and carefully performed psychological experiment can provide sufficient support to a particular model without transformation (an FM principle). (2) Validation of a particular model must precede any kind of numerical estimation, because the estimates depend on the nature of the model, and their statistical properties are conditional upon the model (a CMT and statistical principle). (3) If a transformation is needed, it should be unique and not scale dependent. The transformation should be based on the nature of the data, but it should not capitalize on chance fluctuations in the data. (4) Expected normal scores tests represent "a brilliant tour de force combining the versatility of distribution free tests with the efficiency of the classical ones" (Bradley, 1968, p. 147). The lower bound of the asymptotic relative efficiency (ARE) for each of these tests, compared with their parametric counterpart, is 1. McSweeney and Penfield (1969) have shown empirically that these tests are more powerful than the regular nonparametric ANOVA under a variety of conditions (see also Marascuilo & McSweeney, 1977).

THEORETICALLY DERIVED TRANSFORMATIONS

A very likely reply to the criticism and suggestions in the previous section, in the spirit of FM, is that the choice of transformation, evaluation of particular empirical results, and validation of specific models should be done in the context of prior theoretical and empirical knowledge. This is the theoretical Objective 3 mentioned earlier. It must be noted in this regard, however, that often not enough prior information exists, and, in addition, that a very large body of research with high internal

consistency can be generated using repeatedly incorrect assumptions and/or techniques. However, this theoretically based approach can work if it is conducted in a well defined and systematic framework, as proposed and demonstrated in the work of Birnbaum (1974) and Birnbaum and Veit (1974) on "scale convergence."

FINAL COMMENTS

The above examples should not be regarded as "typical" situations occurring in information integration research, but rather as illustrations of a larger class of possible problems that may arise when using ANOVA to validate models. These particular examples were chosen because of their simplicity and clarity. The frequency of such cases with real data, as well as the number of times these problems have not been recognized in the past, is unknown.

Some final words related to the problem of transformations must be added. We made a distinction that does not exist in the original FM formulation between the original scale and any other "arbitrary" numerical transformation of it. There is enough experimental evidence (Anderson, 1974) to show that the former can be carefully measured so that artificial effects (such as floor or ceiling) are eliminated and so that its units have an intuitive meaning for the experimenter as well as for the subjects. This is not necessarily the case with the other "arbitrary" scales derived by the numerical transformations, and the very least one would want to assume is that the rescaling procedure is uniquely determined and that the transformed scores are suitable for the analysis applied. The proposed algorithm satisfies these requirements.

Yet to be established is the precise relation between the power and efficiency of the proposed procedure and size of the sample and design. Clearly, as sample size increases, the standard errors of estimates of the means and variances decrease, and as design size increases, the number of points on which the estimated mean-variance relation is based increases. Therefore, the reliability of the mean-variance relation is directly related to both sample size and design size.

As sample size increases, independently of design size, the efficiency of the transformations to normal or random normal scores increases, since the distribution of the scores approaches a standard normal. As design size increases, the maximal fit transformation, used in the scale estimation stage, becomes more tightly constrained, and therefore more meaningful. Increased sample size is also beneficial here since it improves the estimates of the individual cell means.

It is quite obvious that the procedure will be efficient for large designs and large samples. However, the precise relation between sample and design size on the one hand, and the power and efficiency of the procedure on the other, is an open empirical question.

REFERENCE NOTES

1. Anderson, N. H. *Algebraic models for information integration* (Tech. Rep. CHIP 45). La Jolla, Calif: Center for Human Information Processing, University of California, San Diego, 1974.

2. Holt, J. O. III, & Wallsten, T. S. *A user's manual for CONJOINT: A computer program for evaluating certain conjoint measurement axioms* (Research Memorandum No. 42). Chapel Hill, N.C: L. L. Thurstone Psychometric Laboratory, University of North Carolina, October 1974.

REFERENCES

ANDERSON, N. H. On the quantification of Miller's conflict theory. *Psychological Review*, 1962, **69**, 400-411.

ANDERSON, N. H. Functional measurement and psychophysical judgment. *Psychological Review*, 1970, **77**, 153-170.

ANDERSON, N. H. Integration theory and attitude change. *Psychological Review*, 1971, **78**, 171-206.

ANDERSON, N. H. Information integration theory: A brief survey. In D. H. Krantz, R. C. Atkinson, R. D. Luce, & P. Suppes (Eds.), *Contemporary developments in mathematical psychology* (Vol. 2). San Francisco: Freeman, 1974.

ANDERSON, N. H. How functional measurement can yield validated interval scales of mental quantities. *Journal of Applied Psychology*, 1976, **61**, 677-692.

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

BARTLETT, M. S. The use of transformations. *Biometrics*, 1947, **3**, 39-52.

BIRNBAUM, M. H. The devil rides again: Correlation as an index of fit. *Psychological Bulletin*, 1973, **79**, 239-242.

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

BIRNBAUM, M. H., & VEIT, C. T. Scale convergence as a criterion for rescaling: Information integration with difference, ratio and averaging tasks. *Perception & Psychophysics*, 1974, **75**, 7-15.

BRADLEY, J. V. *Distribution free statistical tests.* Englewood Cliffs: N.J: Prentice-Hall, 1968.

CURTIS, J. H. On transformations used in the analysis of variance. *Annals of Mathematical Statistics*, 1943, **14**, 107-122.

GLASS, G. V., PECKHAM, P. D., & SANDERS, J. R. Consequences of failure to meet assumptions underlying the analysis of variance and covariance. *Review of Educational Research*, 1972, **42**, 237-288.

KRANTZ, D. H., & TVERSKY, A. Conjoint measurement analysis of composition rules in psychology. *Psychological Review*, 1971, **78**, 151-169.

MARASCUILO, L. A., & MCSWEENEY, M. *Nonparametric and distribution-free methods for the social sciences.* Belmont, Calif: Wadsworth, 1977.

MCSWEENEY, M., & PENFIELD, D. The normal scores test for the C sample problem. *British Journal of Mathematical and Statistical Psychology*, 1969, **22**, 177-192.

RAPOPORT, A., & WALLSTEN, T. S. Individual decision behavior. *Annual Review of Psychology*, 1972, **23**, 131-176.

SMITH, J. E. K. Data transformations in analysis of variance. *Journal of Verbal Learning and Verbal Behavior*, 1976, **15**, 339-346.

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

(Received for publication July 25, 1979.)