Perception & Psychophysics
1989, 45 (3), 275-278

Distributional versus error-filled procedures for transformation

MICHAEL H. BIRNBAUM
California State University, Fullerton, California

The comment by Carterette and Anderson (1987) on Birnbaum (1982b) is misleading because their definitions for three key terms—*distribution*, *replication*, and *scale-free*—differ from the definitions used in Birnbaum's (1982b) paper. By clarifying differences in definition from issues of real disagreement, this note attempts to clarify the fundamental difference between transformation methods used by Birnbaum and Veit (1974), Birnbaum and Elmasian (1977), and Carterette and Anderson (1979), on the one hand, and the very different, distributional procedure discussed by Birnbaum (1982b). To understand the issues involved in transformation techniques, it is useful to review the experiment and procedures of Birnbaum and Elmasian (1977).

Transformation of Repetitions

Birnbaum and Elmasian (1977) asked subjects to judge ratios and differences of loudness of 45 tone pairs, constructed from a 5 × 9, first tone × second tone factorial design. The entire set was repeated 10 times in each session, and each subject served in four sessions, two for each task. Each of four sets of 450 judgments for each subject was separately transformed to fit the following model:

$$J_T^{-1}(T'_{ijk}) = b_j - a_i + e_k, \quad (1)$$

where J_T^{-1} is the strictly monotonic transformation for Task T; T'_{ijk} is the response to the kth repetition of stimulus pair ij with scale values b_j and a_i; and e_k is the effect of repetition block (see Birnbaum & Elmasian, 1977, p. 387, Equation 5). Birnbaum and Veit (1974) had previously treated repetitions and other counterbalanced position effects by a similar procedure.

Carterette and Anderson (1979) treated repetitions (which they called "replications") by means of a variation on Equation 1, although they never explicitly stated the constraints imposed on their transformations.[1] Their novel argument (and an issue that was disputed by Birnbaum, 1982b) was that standard analysis of variance (ANOVA) could be applied to the transformed scores.[2] Birnbaum's (1982b) main purpose was to question whether such procedures really utilize distributional information, as had been argued by Carterette and Anderson (1979), and to discuss the consequences of a different procedure that does impose distributional constraints.

Correspondence may be sent to the author at Department of Psychology, California State University, Fullerton, CA 92634.

As Birnbaum (1982b) noted, the application of ANOVA to the transformed scores presumes that *after* transformation, the distributions of errors are homogeneous and normally distributed (see also Busemeyer, 1980). In a normal distribution, the mean equals the median. The median of the transformed scores containing repetition errors is the same as the single monotonic transformation of the medians. Therefore, a procedure involving transformation of several error-filled repetitions places no additional theoretical constraints, in principle, above and beyond transformation of the medians. Furthermore, since most judgment data also have the property that means and medians are monotonically related, the same results should be obtained with transformation of the means as medians.

However, Carterette and Anderson (1979) stated that their difference judgments were not parallel when transformed by one method, but were parallel when transformed by the other. This result suggests either that their computer program had not found the global best solution, that the use of separate transformations introduced a systematic bias, or that the assumptions of independence of the transformed repeated measures and/or symmetry of the distributions were severely violated, thereby vitiating the use of the F distribution in the manner used.

The actual transformation procedures of Birnbaum and Elmasian (1977) and of Carterette and Anderson (1979) were quite similar: the biggest difference seems to be that Birnbaum and Elmasian (1977) had 10 repetitions per transformation, with a separate transformation for each session, whereas Carterette and Anderson (1979) had five repetitions per transformation, with a separate transformation for each half-session. Neither of those *procedures* is based on any theory of the distribution of errors.

Fitting Distributions versus Transforming Errors

Birnbaum's (1982b) distributional method is very different from the above procedures. Carterette and Anderson's (1987) description of Birnbaum's (1982b) suggestion as "transforming several replications at a time" would have been a correct description of the method used by Birnbaum and Veit (1974) or Birnbaum and Elmasian (1977); however, it is not correct as a description of Birnbaum's (1982b) analysis. Perhaps because Carterette and Anderson's (1987) use of the term "distribution" differed from that of Birnbaum (1982b), they failed to appreciate the difference between the two procedures.

Birnbaum's (1982b) suggestion was that rather than transforming several repetitions (or "replications"), one can transform an estimate of the distribution function itself. In other words, the dependent variable is an estimate of the cumulative probability of a score falling below a given value, rather than the score itself. In this procedure, the transformation is interpreted as the inverse of the distribution function and is *not* an estimate of the judgment function, as it would be using methods as in

Equation 1. In Birnbaum's (1982b) usage, a *distributional* method is one in which a theory of the distribution function is incorporated in the transformation procedure, making transparent the vulnerability of the procedure to proper theory.

To illustrate a distributional method, let P_{ijm} be the proportion of judgments less than T'_m for stimulus combination ij. One possible theory is that the distributions are identical except for their central tendency (Birnbaum, 1982b, Equation 4). This transformation theory can be rewritten as follows:

$$F^{-1}[P(T'_{ijk} < T'_m)] = \Psi_{ij} - t_m, \quad (2)$$

where t_m is the estimated value of response T'_m and is interpreted as $J^{-1}(T'_m)$; F is the distribution function; and Ψ_{ij} is the subjective value of the combination of stimulus i and j. Note that this transformation procedure assumes that the distribution function, F, is the same for all stimulus combinations, which *may or may not* be compatible with any particular theory of Ψ_{ij}. For example, the data may or may not be compatible with additivity:

$$F^{-1}[P(T'_{ijk} < T'_m)] = a_i + b_j - t_m, \quad (3)$$

where $\Psi_{ij} = a_i + b_j$.

Table 1 shows hypothetical data that would yield different conclusions when transformed by the error-filled methods (Equation 1) or the particular (homogeneous) distribution theory of Equation 2. The method of Equation 1 concludes that the data of Table 1 are additive, because the log of $(T' - .75)$ renders the medians additive. The method of Equation 2, however, implies that if the distributions are assumed to be homogeneous, then the data are not additive. Note that Table 1 violates joint independence (Krantz & Tversky, 1971); therefore, Equation 3 is not compatible with the data. In other words, the homogeneity assumption that drives the transformation of Equation 2 is not compatible with the additive combination of Factors A and B.

It would also be possible to generate a case in which the medians are an additive combination of A and B before transformation, but Equation 2 would require that A and B are not additive. To create such a case, let the spread of the distributions be different for different means.

Table 1
Values of $P(T'_{ijk} < T'_m)$ for Example Data

Stimulus Combination	Values of T'_m										
	.5	1	1.5	2	2.5	3	3.5	4	4.5	5	5.5
A_1B_1	.27	.38	.50	62	.73	.82	.88	.92	.95	.97	.98
A_1B_2	.18	.27	.38	.50	.62	.73	.82	.88	.92	.95	.97
A_2B_1	.08	.12	.18	.27	.38	.50	.62	.73	.82	.88	.92
A_2B_2	.02	03	.05	.08	.12	.18	.27	.38	.50	.62	.73

Note—The transformation $M(T') = \log(T' - .75)$ renders the medians additive; however, that transformation would violate the assumption of Equation 2, which implies that the T'_m are evenly spaced and that A and B are not additive. This example shows that additivity and homogeneity need not be compatible. This example was constructed from the equation, $P = 1/[1 + \exp(M - T')]$, where $M = 1.5$, 2, 3, and 4.5, for the four rows, respectively.

Because Birnbaum's (1982b) procedure can lead to conclusions different from those of the procedures used by Birnbaum and Elmasian (1977) or Carterette and Anderson (1979) (as it does in Table 1), it should not be confused with those procedures. Carterette and Anderson (1987) imply that they employed Birnbaum's (1982b) procedure; however, it is important to be clear that, while they did use a procedure similar to that of Birnbaum and Elmasian (1977), and they did use the term "distributional" to describe it, they did not use the procedure of Birnbaum (1982b), which involves the distribution function itself, as in Equations 2 and 3.

Birnbaum's (1982b) article did not advocate the distributional method so much as to present an exposition of what the concept of a distributional method entails. Homogeneity of distributions would constitute only one of many possible distributional procedures for transformation, but as Birnbaum's example (1982b, Tables 2 and 3) illustrates, there is an indeterminancy between distribution theory and the algebraic model under investigation (see also Eisler, 1965). Therefore, the distributional methods should be regarded as theories of transformation rather than theory-neutral algorithms.

Scale-free Tests and Real Disagreements

Although Carterette and Anderson (1979) used the term "scale-free" in their article, their definition differs from that of Birnbaum and Veit (1974). The scale-free test was designed by Birnbaum (1974) and Birnbaum and Veit (1974) to distinguish between models that are not ordinally distinguishable in the usual experiment. For example, the additive model can be transformed to a multiplicative model by exponential transformation. The scale-free test imposes additional ordinal constraints that permit members of the additive family to be distinguished (Birnbaum, 1982a). The bisection task of Carterette and Anderson (1979) would not qualify as a scale-free test of the bisection model, according to Birnbaum and Veit (1974), because the ordinal properties of the data would not, in principle, permit one to test the parallel bisection model against the geometric (bilinear) bisection model.

Even more importantly, the same data have received different theoretical interpretations when analyzed by Anderson's methods and Birnbaum's. These real and extremely important controversies deserve fuller exploration and should not be lost among confusions that arise from different usages of terms. For example, Anderson (1983) cited two experiments on equity and inequity as illustrations of his "two-operation logic," and argued that the scale-free test is just a form of this two-operation logic. Mellers (1985) reanalyzed Anderson's data, using scale-free methods, as in Birnbaum (1974) and Birnbaum and Veit (1974). Mellers found that the scale-free approach led to different conclusions when applied to Anderson's

data. In both cases, the two-operation logic led to conclusions that were not compatible with the data, when the data were reanalyzed. As Mellers (1985, p. 514) put it, Anderson's (1983) two-operation logic may "lead to illogical conclusions and should not necessarily be referred to as a logical procedure."

Birnbaum (1982a) has also analyzed other examples in which the different approaches led to different conclusions. The most relevant example for the present discussion is the scale-free test of the additive or parallel-averaging models of the size-weight illusion (Birnbaum & Veit, 1974). Birnbaum and Veit found that the judged difference in heaviness between two different-sized blocks of the same weight was greater when the (common) weight was greater. In a second experiment, subjects compared a series of size-weight combinations to several standards, and the subtractive model was used as the criterion for transformation. Birnbaum and Veit concluded that if difference judgments are represented either by a subtractive or ratio model, the additive or parallel-averaging models must be rejected for the size-weight illusion. Anderson's (1977, 1981, 1983) reviews of this topic reached conclusions (based on scale-dependent research) different from those of Birnbaum and Veit (1974), although he did not cite their study. Thus, despite a similarity of purpose and terminology, Anderson's (1983) two-operation logic should not be confused with the scale-free test as used by Birnbaum (1974) and by Birnbaum and Veit (1974).

There also is an empirical difference between Birnbaum and Elmasian (1977) and Carterette and Anderson (1979) that deserves further comment. Birnbaum and Elmasian (1977) concluded that loudness difference judgments can be represented by the subtractive model of comparison. Carterette and Anderson (1979) used different experimental procedures (noises rather than 1000-Hz tones, shorter interstimulus intervals, etc.), and found a systematic discrepancy in their difference judgments that could be transformed away by one procedure but not by the other. Carterette and Anderson (1979) were willing to assert that by using independent transformations for different replicates (of five repetitions), they had a valid statistical theory, and although they found that the different methods of transformation led to different conclusions for their data, they preferred one method and concluded that the subtractive model was systematically violated. Although they concluded that there was a real discrepancy, they did not determine its origin. Among the possible interpretations is the idea that with short interstimulus intervals, the loudness of the first noise affects the loudness of the second, which might appear as a discrepancy from any model that assumes that the stimuli do not affect one another. Another possibility is that the response scale was nonlinear, but Carterette and Anderson's (1979) transformation procedures failed to work properly, leading them to an erroneous conclusion.

REFERENCES

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

ANDERSON, N. H. (1981). *Foundations of information integration theory*. New York: Academic Press.

ANDERSON, N. H. (1983). Ratio models of equity and inequity: Comment on Mellers. *Journal of Experimental Psychology: General*, **112**, 513-515.

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

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

BIRNBAUM, M. H. (1982b) On rescaling data to fit the model and concluding that the model fits: A note on monotone transformation. *Perception & Psychophysics*, **32**, 293-296.

BIRNBAUM, M. H., & ELMASIAN, R. (1977). Loudness "ratios" and "differences" involve the same psychophysical operation. *Perception & Psychophysics*, **22**, 383-391.

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

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

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

CARTERETTE, E. C., & ANDERSON, N. H. (1987). Setting straight the record. *Perception & Psychophysics*, **42**, 409-410.

EISLER, H. (1965). The connection between magnitude and discrimination scales and direct and indirect scaling methods. *Psychometrika*, **30**, 271-289

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

MELLERS, B A (1985) A reconsideration of two-person inequity judgments A reply to Anderson. *Journal of Experimental Psychology: General*, **114**, 514-520.

NOTES

1. In many judgment experiments, subjects make repeated judgments of the entire set of stimuli, which are randomized and presented several times. Sometimes, these repeated measures are loosely called *replicates*. However, human subjects who give repeated measures are learning in the course of the experiment; their responses are correlated, and they can change systematically during the course of an experiment. The term *replications* refers to independent measures, obtained under identical conditions that are interchangeable; *repetitions* are repeated measures that may be correlated, may show order effects, and are not interchangeable.

Carterette and Anderson (1987) use the term "replication," where Birnbaum and Elmasian (1977) had used the term "repetition," and some of their difficulty with Birnbaum's (1982b) article can perhaps be traced to confusion arising from different usages of that term. This difference in definitions may also explain why their 1979 paper incorrectly described the procedure of Birnbaum and Elmasian (1977).

2 Their ANOVA argument was based on the use of separate monotone transformations applied to subsets of the data. "Two independent estimates of the best monotone transformation were thus obtained for each subject" (Carterette & Anderson, 1979, p. 275). The two so-called "independent estimates," if they were truly independent, might justify the term "replicate," and that is why Birnbaum described their procedure as using a separate transformation for each replicate. This interpretation is also consistent with Anderson's (1977) statement:

> Fortunately, there is a simple way around these problems The monotone transformation program is applied separately to each replication of the de-

> sign. Since the transformed data are then independent across replications, any one degree of freedom component of the interaction has a valid test. (p. 210)

Although Anderson's (1977) statement could be interpreted as a definition of replication, Carterette and Anderson (1987) state that they "severely criticized" the use of one replication per transformation, in disagreement with Anderson (1977).

In any case, confusion over the terms "replicate" or "repetition" is beside the point, because Birnbaum's (1982b) paper challenged whether the statistical arguments of Carterette and Anderson (1979) were justified by the use of independent transformations, whether or not several repetitions are involved in each replication. Birnbaum (1982b) did not mean to imply that Carterette and Anderson (1979) did not have several repetitions in each transformation, a procedure common to Birnbaum and Veit (1974), Birnbaum and Elmasian (1977), and Carterette and Anderson (1979).

(Manuscript received December 4, 1987;
revision accepted for publication August 24, 1988.)