Psychological Bulletin
1974, Vol. 81, No. 11, 854–859

REPLY TO THE DEVIL'S ADVOCATES: DON'T CONFOUND MODEL TESTING AND MEASUREMENT

MICHAEL H. BIRNBAUM [1]

Kansas State University

Replies by E. F. Alf and N. M. Abrahams and by L. G. Rorer defend the correlation-regression approach to model testing, contending that if a priori measurements are assumed to be proper psychological values and if the correct model is known, correlations can be higher for the better model. But since psychologists cannot know in advance the correct scales and models, popular correlational techniques are inappropriate for investigating psychological processes. It is necessary to separate measurement from the evaluation of a model. A further attempt is made here to clarify the relationships between different methods of analysis.

Birnbaum (1973) criticized a currently popular use of correlation that confounds measurement with model testing, demonstrating that a poorer model can achieve higher correlations with the data when a priori measurements are used.

Recent replies by Alf and Abrahams (1974) and by Rorer (1974) correctly contend that once the data have been properly diagnosed by other techniques, it may be possible to use regression so that the correlation coefficient is higher for the correct model. But the fundamental question should be: What are the advantages or disadvantages of correlational techniques for exploring psychological theories under conditions where the correct models and psychological values of the stimuli are unknown?

Under these conditions, correlations of fit can be misleading since they depend on such factors as (*a*) unreliability of response, (*b*) experimental design (which includes variation and covariation of independent variables), (*c*) stimulus metric, (*d*) response metric, and (*e*) number of estimated parameters, as well as (*f*) the "goodness" of the model. When correlational analyses are reported, the journal reader has no way of knowing what the original data (and the pattern of deviations of fit) look like. Consequently, unless the data are appropriately presented, the correlation coefficient cannot be unambiguously interpreted as an index of fit (Anderson, 1971, 1972; Birnbaum, 1973; Darlington, 1968).

Alf and Abrahams (1974) and Rorer (1974) argued that correlations measure the "goodness" of the model, basing their arguments on the assumption that the evaluation of psychological laws is logically dependent on a priori measurements of the variables. But the appropriate scaling of psychological variables can hardly be known a priori. The present paper is a further effort to discuss an approach that separates measurement from model testing by scaling the stimuli in accord with the model to be tested (Anderson, 1970, 1971, 1972; Krantz, Luce, Suppes, & Tversky, 1971; Krantz & Tversky, 1971).

Measurement and Model Testing

Measurement and model testing go hand in hand. The basic idea of measurement is to assign numerical values to objects so that laws relating the measured variables describe empirical relationships among the objects. The basic idea of model testing is to ask whether a particular set of assumptions can account for a set of observed relationships. Measurements "make sense" with respect to empirical laws relating theoretical variables that can be measured.

Hopefully, the world is simple and can be described by a set of laws that are interlocked by a small number of measured varia-

[1] The author thanks Allen Parducci, Norman H. Anderson, Clairice T. Veit, and James C. Shanteau for helpful suggestions.

Requests for reprints should be sent to Michael H. Birnbaum, who is now at the Department of Psychology, University of Illinois at Urbana-Champaign, Champaign, Illinois 61820.

bles. Measurements of the same variable derived from different empirical laws should agree. In physics, the agreement is so good that it is often taken for granted that measurements of length, time, and mass, for example, can be measured independently of the model under investigation. But even in physics the measurement approach has its advantages (see Footnote 3).

Figure 1 presents an outline of psychological measurement for a situation in which the response depends on the combined effect of two variables. Extension to a greater number of variables is straightforward. The stimuli, indexed i and j, have subjective values s_i and s_j. The integration function, I, represents the psychological law, or model, that describes how the subjective values combine to form an overall impression, Ψ_{ij}. The overt response, R_{ij}, is assumed to be a monotonic function, J, of the psychological impressions.

There are three psychological issues to be distinguished: (a) finding the subjective values, s_i and s_j, is called *measurement* (scaling); (b) establishing and *testing the model,* $\Psi_{ij} = I(s_i, s_j)$; (c) finding the *response function,* J, between the integrated impression, Ψ, and overt response, R.

The measurement approach derives scales (s) from the data in accord with the model to be tested. In principle, one need not have a priori measurements of the stimuli. In the special case where simple physical measurements are available, the relationship between subjective values and physical values, ϕ, is called the *psychophysical function,* $s = H(\phi)$. The form of H can be determined after fitting the model and need not be assumed to test the model. Thus, psychophysical scaling can be separated from the test of the model.

The popular regression approach attempts to fit the data as some function of the a priori measurements of the stimuli, $R = F(\phi_i, \phi_j)$. This function is a composition that confounds the stimulus scale, H, response scale, J, and theoretical model, I, all in one. Every different possible transformation between a priori values and subjective values, H, results in a different F and is therefore considered a different "model." A user of these techniques could easily find himself computing hundreds of correlations of fit for different equations that represent the same I function. Omission of one "model" could be disastrous, since this procedure requires one to fit every possible equation and choose the one with the highest correlation. Even worse, when subjective values cannot be expressed as a function of a priori values (i.e., if two stimuli with the same a priori value have different psychological values), the approach is doomed to failure.

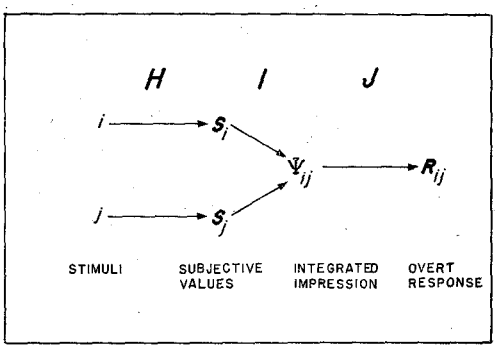

FIGURE 1. Outline of psychological measurement. (The stimuli, referenced by the indexes i and j, have subjective values s_i and s_j; they combine according to the psychological law, $\Psi = I(s_i, s_j)$, to form an integrated impression, Ψ_{ij}, which is related to the overt response, R, by the response function, $R = J(\Psi)$.)

The fact that F is a confounded composition implies that correlations of fit for any function, F, cannot be unambiguously interpreted. Very little can be inferred from the finding that one equation correlates higher with the data than another; another possible equation excluded from the analysis might have correlated even higher.

The measurement approach permits an evaluation of the possibilities for the psychological model, I, without making any assumptions about H and assuming only that J is monotonic. Each psychological issue can be pursued separately. Separate evaluation of H, I, and J enhances understanding without sacrificing prediction. The following examples illustrate how models lead to measurements and show that a priori measurements are unnecessary and often inappropriate for the evaluation of a psychological law.

Examples of Psychological Measurement

Additive model. Figure 2 plots the hypothetical data of Birnbaum (1973, Table 2) against subjective values of the stimuli derived from the additive model, $\Psi_{ij} = s_i + s_j$. When J is linear, the additive model predicts that the curves will be parallel. The curves *need not* be linear functions of the a priori values of the stimuli. Since the curves in Figure 2A are parallel, the marginal means are linearly related to the additive model scale values (Anderson, 1970, 1971). The figure shows that the subjective values are equally spaced. In Birnbaum (1973, Figure 2B), the same data are nonlinear when plotted against a priori values of 1, 2, 4, 8, and 16. In this example, the subjective values are logarithmically related to the a priori values.

The correlation approach advocated by Alf and Abrahams (1974) and Rorer (1974) confuses the linearity of the curves (the form of H) with their parallelism (the form of I). Consequently, although the additive model correlates perfectly with the data when the subjective scales are employed, the correlation using a priori stimulus values is .933, less than the .998 correlation for the logarithmic form of the multiplicative model. The contention that the higher "correlation truly reflects the superiority of the multiplicative model [Alf & Abrahams, 1974, p. 73]" is misleading since the multiplicative model predicts a diverging fan of curves.

The additive model can be tested very simply, without the use of either a priori measurements, correlation coefficients, or any elaborate statistics. All one need do is plot the data and visually inspect whether the curves are parallel. If the curves are parallel, an additive representation of the data exists. If the curves are systematically nonparallel, further analyses (discussed below) may be required to assess the possibility and propriety of transforming the data to parallelism.

There are regression techniques for fitting the additive model without relying on a priori measurements. An appropriate technique would employ dummy variables for the levels of the independent variables or use polynomial expansion of the a priori scales (see

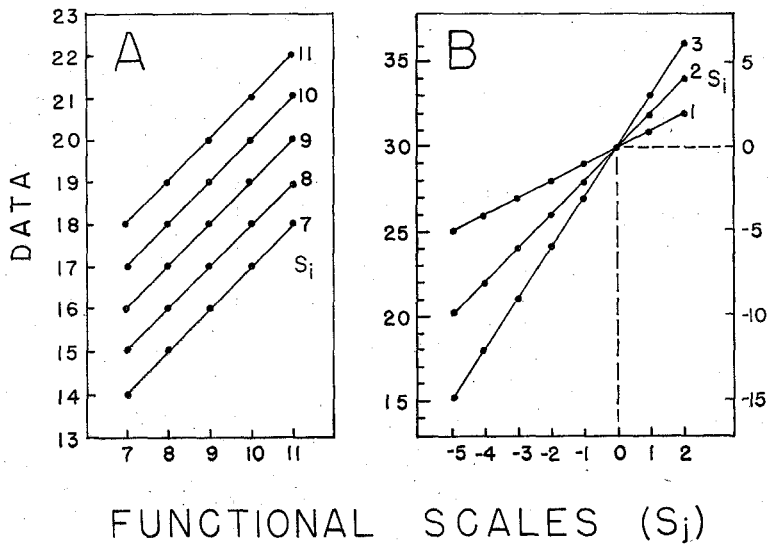

FIGURE 2. A: Hypothetical additive data. (Scales derived in accord with the additive model $\Psi_{ij} = s_i + s_j$, will reproduce each data point. Curves are linear functions of subjective scales, but would be logarithmically related to a priori scales; see text.) B: Hypothetical multiplicative data. (Scales derived from the multiplicative model, $\Psi_{ij} = s_i \cdot s_j$, reproduce each data point. Abscissa values are spaced so that curves form a bilinear fan, determining scale values; slopes are scale values for other variables. Intersection of curves defines the location of the zero points of the scales.)

e.g., Cohen, 1968). When combined with appropriate tests of deviations of fit, such techniques may be useful in instances where complete factorial designs are not feasible (as illustrated in Birnbaum, 1974b); however, nonorthogonality of the independent variables can create additional problems for correlations (Darlington, 1968).[2]

Multiplicative model. Figure 2B plots the data from Table 2 of Birnbaum (1973). The crossing curves clearly violate the additive model. The functional scales (Anderson, 1970, 1971) have been derived from the data in accord with the multiplicative model, $\Psi_{ij} = s_i \cdot s_j$. The projections of the crossover correspond to the functional zero points for the dependent variable and the independent variable plotted on the abscissa.[3]

Birnbaum (1973) showed that if the subjective values are linearly related to a priori values, currently popular correlational techniques could indicate that the additive model provides a "better" representation of these data, correlating .930 (compared with .899) in spite of horrendous deviations.[4]

Data transformation. The third example, for a symmetric, factorial design (Figure 3A), shows a divergent interaction that would be inconsistent with both additive and multiplicative models if the response function, J, were assumed to be linear. Such an interaction may be attributed to nonlinearity of J rather than nonadditivity of the integration function. If the dependent variable, R, is considered to be only an ordinal measure of the psychological variable, then "additivity" has a weaker, more general definition (see Krantz et al., 1971; Krantz & Tversky, 1971). A data matrix would be termed "additive" if it were possible to find scales s_i and s_j, such that $R_{ij} > R_{kl}$ whenever $s_i + s_j > s_k + s_l$; that is, if scales can be found so that the additive model correctly generates the empirical ordering. The data of Figure 3A are "additive" in this ordinal sense, as shown in Figure 3B where the dependent variable has undergone a square-root transformation. As can be seen, the transformation (interpreted as J^{-1}) renders the curves parallel. Once the curves are parallel, the marginal means again estimate the scale values of the stimuli, which in this case are nonlinearly related to the a priori values, but linearly related to 1, 5, 8, 10, and 11.

A difference between functional measurement (Anderson, 1970, 1971) and conjoint measurement (Krantz et al., 1971; Krantz & Tversky, 1971) has centered over the propriety of rescaling data that would be incon-

[2] Correlations of fit are highly dependent on the experimental design. Dudycha and Naylor (1966) compared correlations and concluded that cue intercorrelations affect the way subjects process information. However, Schenk and Naylor (1968) cautioned that the results may be a statistical artifact of the correlation coefficient. Birnbaum and Veit (1973) and Birnbaum, Kobernick, and Veit (1974) illustrated more appropriate techniques that unconfound the experimental design from the statistical analysis.

[3] The ideal gas law, that the volume of gas is directly proportional to the ratio of temperature to pressure, could be used to measure temperature and determine absolute zero. But if a priori measures of temperature (Centigrade or Fahrenheit) were used, an investigator comparing correlation coefficients might erroneously have concluded that temperature and pressure combine additively unless he made the appropriate graph of the data. Plotting volume as a function of temperature with a separate curve for each level of pressure would reveal a bilinear fan of curves that intersect near $-273°$ Centigrade.

[4] Rorer (1974) implicitly assumed that the zero points of the a priori scales are the subjective zero points. Unfortunately, Rorer made several additional errors that could lead to confusion unless corrected: (*a*) In his analysis of the multiplicative model, Rorer must have meant "power function" ($Y = X^a$) rather than "exponential" ($Y = a^X$). Even with this correction, his analysis remains in error because (*b*) he confused logarithmic transformation with addition of a priori cross-products to a linear equation, and misreported Birnbaum's (1973) analysis. (*c*) Rorer applied the popular (but erroneous) interpretation of the magnitude of the linear coefficients in the equation $Y = 4\ X_1 + 6\ X_2 - X_1 \cdot X_2 + 6$ to imply that one should expect the linear model to correlate highly with data generated from the equation. The coefficients cannot be interpreted so simply. For example, the linear model would achieve a correlation of zero with data generated using a 3×3 factorial design with levels of $X_1 = 5$, 4, and 3, and levels of $X_2 = 7$, 6, and 5. (*d*) He confused the psychophysical function with nonlinear transformation of the dependent variable. (*e*) He confused parallelism of the curves with absence of a significant X_1X_2 term. Even if the Xs were properly scaled variables, absence of bilinear interaction does not imply additivity. Rorer is correct, however, in his assertion that a comparison of correlations cannot be unambiguously interpreted.

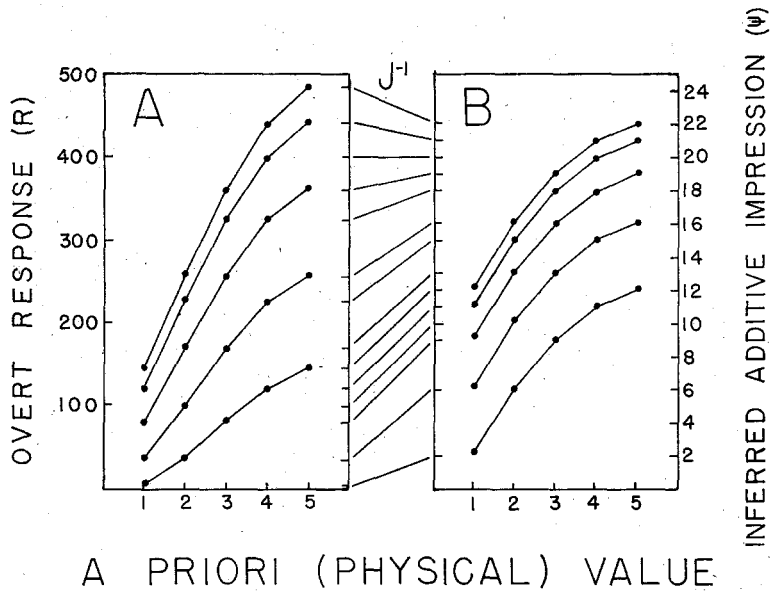

FIGURE 3. Hypothetical data to illustrate data transformation. A: The nonparallelism of the curves is inconsistent with the additive model under the assumption that the J function is linear. B: Each point is the square root of the corresponding point in Panel A. (Parallelism indicates that the transformed data are additive; nonlinearity of the curves merely indicates that the subjective values are a negatively accelerated function of the physical values. Transformation may be theoretically inappropriate in certain circumstances; see text.)

sistent with the model unless transformed. Functional measurement allows rescaling but has resisted transformation of data obtained under experimental conditions where simple models have previously fit without rescaling. Conjoint measurement has accepted ordinal violations of a model as convincing, arguing that deviations that can be removed by monotone transformation may be without psychological significance (Krantz et al., 1971). Birnbaum (1974a) discussed criteria for appropriate transformation (see also Birnbaum & Veit, 1974) and applied scale-free techniques that can determine whether an interaction such as that shown in Figure 3A is "real," or should be transformed as in Figure 3B.

Other Problems with Correlations

Aside from the measurement problem, two well-known criticisms of correlations deserve mention: (*a*) high correlations do not insure a good fit. They can easily coexist with serious model discrepancies (Anderson, 1971, 1972). For example, the data in Figure 2B correlate .930 with the additive model. Proper evaluation of a model must attend to discrepancies. Plots of predicted versus obtained are not generally adequate to portray and assess discrepancies of fit. (*b*) The squared multiple correlation generally represents the ratio of predicted to total variance. Since the investigator controls the experimental design and hence the total variance, he also controls the magnitude of the correlation coefficient. Furthermore, when correlations of fit are compared, certain experimental designs favor one model over another. Consequently, correlation is neither an "absolute" index such that a certain value could be considered a "good" fit, nor is it a "relative" index such that correlations for different models can be unambiguously compared.

Tell Truth and Shame the Devil

Let's *give* the devil his due: The devil *can* speak true, *sometimes*. *Sometimes* correlation means causation; *sometimes* a high correla-

tion of fit means that a model is a good representation of a psychological process; *sometimes* the "better" representation correlates higher. Although correlation can be used appropriately, in its current uses it is the tool of the devil.

When misused, the tools of statistics often deceive, conveying misimpressions of the original data from which they were calculated (Huff, 1954). The cautious reader should think of Figure 2B next time he sees a correlation in the .90s. Unless the data were appropriately presented, the reader would have no way of knowing that supposedly additive data (.93 correlation) contain critical violations of the theory.

Shakespeare wrote, "I can teach thee, coz, to shame the devil by telling truth." To tell the truth, psychologists should report graphs of the data that allow inspection of critical predictions of the theory. It is less important to know that the fit is "pretty good" than it is to know that the deviations are *not* bad.

REFERENCES

Alf, E. F., & Abrahams, N. M. Let's give the devil his due: A response to Birnbaum. *Psychological Bulletin,* 1974, **81,** 72–73.

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. Looking for configurality in clinical judgment. *Psychological Bulletin,* 1972, **78,** 93–102.

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. (a)

Birnbaum, M. H. Using contextual effects to derive psychophysical scales. *Perception & Psychophysics,* 1974, **15,** 89–96.

Birnbaum, M. H., Kobernick, M., & Veit, C. T. Subjective correlation and the size-numerosity illusion. *Journal of Experimetal Psychology,* 1974, **102,** 537–539.

Birnbaum, M. H., & Veit, C. T. Judgmental illusion produced by contrast with expectancy. *Perception & Psychophysics,* 1973, **13,** 149–152.

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,** 7–15.

Cohen, J. Multiple regression as a general data-analytic system. *Psychological Bulletin,* 1968, **70,** 426–443.

Darlington, R. B. Multiple regression in psychological research and practice. *Psychological Bulletin,* 1968, **69,** 161–182.

Dudycha, A. L., & Naylor, J. C. The effect of variations in the cue R matrix upon the obtained policy equation of judges. *Educational and Psychological Measurement,* 1966, **26,** 583–603.

Huff, D. *How to lie with statistics.* New York: Norton, 1954.

Krantz, D. H., Luce, R. D., Suppes, P., & Tversky, A. *Foundations of measurement.* New York: Academic Press, 1971.

Krantz, D. H., & Tversky, A. Conjoint measurement analysis of composition rules in psychology. *Psychological Review,* 1971, **78,** 151–169.

Rorer, L. G. "What, Can the devil speake true?" *Psychological Bulletin,* 1974, **81,** 355–357.

Schenck, E. A., & Naylor, J. C. A cautionary note concerning the use of regression analysis for capturing the strategies of people. *Educational and Psychological Measurement,* 1968, **28,** 3–7.

(Received January 7, 1974)