Alchemy in the Behavioral Sciences
Author(s): Hillel J. Einhorn
Source: *The Public Opinion Quarterly*, Vol. 36, No. 3 (Autumn, 1972), pp. 367-378
Published by: Oxford University Press on behalf of the American Association for Public Opinion Research
Stable URL: http://www.jstor.org/stable/2747445
Accessed: 18/09/2013 04:00

ALCHEMY IN THE BEHAVIORAL SCIENCES*

BY HILLEL J. EINHORN

Access to powerful new computers has encouraged routine use of highly complex analytic techniques, often in the absence of any theory, hypotheses, or model to guide the researcher's expectations of results. The author examines the potential of such techniques for generating spurious results, and urges that in exploratory work the outcome be subjected to a more rigorous criterion than the usual tests of statistical significance.

Hillel Einhorn is Assistant Professor of Behavioral Science, Graduate School of Business, University of Chicago.

WITH THE LARGE-SCALE use of electronic computers, powerful new methods for data analysis have become quite prevalent. Although this development may be viewed with considerable enthusiasm by some,[1] others may view the gains to be derived from increased ability to handle large amounts of data with increasingly sophicated tools with more than a certain degree of skepticism. Such skepticism is based on the observation that as methods and techniques get more complicated, the role of theory in research is being dangerously ignored in favor of purely empirical work that proceeds without so much as a hypothesis. Like Pirandello's characters in search of an author, many of today's researchers seem to have an assortment of techniques in search of a substantive problem.

The general question of proceeding inductively or deductively in science is not easily answered. As Armstrong has put it,

> . . . there is still a great deal of controversy over the relevant roles of theorizing and of empirical analysis. We should note that the problem extends beyond one of scientific methodology; it is also an emotional problem with scientists. There is probably no one reading this paper who is not aware of the proper relationship between theorizing and empirical analysis. On the other hand, we all know of *others* who do not understand the problem.[2]

The basic position of this article is that proceeding via a "dustbowl" empiricism is dangerous at worst and foolish at best. Without some

* This work was supported by a grant from the Spencer Foundation.

[1] B. F. Green, "The Computer Revolution in Psychometrics," *Psychometrika,* Vol. 31, 1966, pp. 437-445.

[2] J. S. Armstrong, "Derivation of Theory by Means of Factor Analysis, or Tom Swift and His Electric Factor Analysis Machine," *American Statistician,* Vol. 17, 1967, p. 17.

set of prior notions, whether full-scale theory, set of hypotheses, or model of some kind, the researcher is without a signpost or benchmark against which to evaluate his results. The purely empirical approach is particularly dangerous in an age when computers and packaged programs are readily available, since there is temptation to substitute immediate empirical analysis for more analytic thought and theory building. It is also probably too much to hope that a majority of researchers will take the time to find out how and why a particular program works. The chief interest will continue to be in the output—the results—with as little delay as possible.

A secondary theme that will become apparent is that if one does engage in purely exploratory research, the results should be subjected to a more rigorous criterion than "statistical significance" or some other statistical criterion. Replication is the backbone of science and when techniques capitalize on chance fluctuations in the particular sample of data at hand, it is imperative to replicate (or cross-validate) the results on a new set of cases. As will become clearer, this is an extremely important consideration that has not always been observed.

What follows is a discussion of several techniques that will illustrate how easy it is to commit Type I errors (rejecting the null hypothesis falsely) when computers are used without much thought. Before beginning, one comment is in order. The techniques to be discussed *are* useful when they are properly applied. The reason for dealing with them here is that they are particularly easy to misuse, with the result that spurious results are taken for real ones. The methods to be discussed include AID (automatic interaction detection), aspects of multiple regression, factor analysis, and nonmetric multidimensional scaling.

AID[3]

One of the most popular of the relatively new and powerful computer-based techniques is the so-called AID (automatic interaction detection) technique.[4] The main purpose of this analysis is ". . . to identify and segregate a set of subgroups which are the best we can find for maximizing our ability to predict the dependent variable."[5] According to Sonquist, "the essence of the algorithm is the sequential

[3] I would like to thank Zvi Lieber for programing the simulation done in this section.

[4] J. N. Morgan and J. A. Sonquist, "Problems in the Analysis of Survey Data and a Proposal," *Journal of the American Statistical Association,* Vol. 58, 1963, pp. 415-435.

[5] *Ibid.,* p. 231.

application of a one-way analysis of variance model."[6] The program searches for the "best" subgroups that can be derived from a set of predictor variables so that these subgroups will serve as the levels on a one-way analysis of variance and effectively explain a maximum amount of variance—maximum in terms of being higher than any other set of subgroups that can be formed from the predictors. The results of an AID analysis resemble a "decision tree," where the total dependent variable is the parent group and the branches of the tree are formed by splitting the parent group according to scores on the various predictor variables. Of particular importance in the analysis is the fact that a linear model is *not* assumed, i.e., the program searches for interactions between the predictors that will best explain the variance of the dependent variable. Although AID was originally developed for use on survey data, it has been used in other areas as well.[7] Recent developments using AID and Multiple Classification Analysis (MCA) are given by Sonquist.[8]

While AID has been suggested as an exploratory device for the generation of hypotheses from data, the technique may be *too* powerful, i.e. it may make sense out of "noise". One can view the technique as similar to a nonlinear regression program which fits a function to data so that the function will go through the mean of each array of the dependent variable given the predictor score. Since no specific functional form is assumed in the analysis, initial fit will usually be quite good although prediction to new samples may be limited.

There are two important issues concerning AID (as well as other computer-based procedures for data analysis): (1) testing hypotheses on the same data that generate those hypotheses, and (2) the use of AID for predictive purposes. With respect to (1), it is poor science to "test" hypotheses on the same data that generated those hypotheses. If AID is used to gain insight and hypotheses from data, these hypotheses can only be tested on *another* set of data. The degree of predictability is a measure of the degree to which the hypotheses are validated. Researchers using AID have not paid enough attention to the question of cross-validation. Where cross-validation has been tried

[6] J. A. Sonquist, "Finding Variables that Work," *Public Opinion Quarterly*, Vol. 33, 1969, pp. 83-95.

[7] J. S. Armstong and J. G. Andress, "Exploratory Analysis of Marketing Data: Trees vs. Regression," *Journal of Marketing Research*, Vol. 7, 1970, pp. 487-492; H. Assael, "Segmenting Markets by Group Purchasing Behavior: An Application of the AID Technique," *Journal of Marketing Research*, Vol. 7, 1970, pp. 153-158; R. Tanofsky, *et al.*, "Pattern Analysis of Biographical Predictors of Success as an Insurance Salesman," *Journal of Applied Psychology*, Vol. 53, 1969, pp. 136-139.

[8] J. A. Sonquist, "Recent Developments in Sequential Data Analysis Strategy," *Proceedings of the Social Statistics Section, American Statistical Association*, 1969, pp. 74-90.

on other configural prediction schemes, large amounts of "shrinkage" have been found.[9]

In order to investigate more fully the fitting of AID to data, the following simulation was performed. Two factors thought to be of special importance were investigated: the sample size and the size of the final subgroups. This latter variable is important since one can increase the amount of explained variance on initial fit by using smaller groups (reducing the number of observations per array—in regression terms—by increasing the number of groups). However, this would also greatly increase the sampling error, so that there should be greater shrinkage on cross-validation. The simulation was set up as follows.[10] A dependent variable was used which consisted of ten levels (1 to 10). Ten independent variables were used with five levels within each predictor. For each simulated observation, a level (value) was assigned to the dependent variable, and similarly, a random procedure was used to assign a level to each of the independent variables. Since sample size was of interest, the simulation was carried out for samples varying from 100 to 1,000 in increments of approximately 100.[11] In order to examine the effects of minimum group size, the simulation was run with the specification that the final group sizes be at least 10, 25, and 50. Linear regressions were also run in order to provide a check on the randomness of the generated data. The amount of variance explained by the two methods was obtained and tested for significance (from zero) by means of the *F*-test. The results are shown in Table 1.

Examination of Table 1 shows that for group sizes of 10, the AID analysis explains a large amount of variance for the smaller sample sizes. For this set of random data, the program does not split the parent group at sample size 700 (a second simulation was performed with a minimum group size of 10 and there was a significant split for

[9] G. A. Forehand and L. L. McQuitty, "Configurations of Factor Standings as Predictors of Educational Achievement," *Educational and Psychological Measurement,* Vol. 19, 1959, pp. 31-43.

[10] The version of the AID program used was developed by Dr. Eli Marks, Wharton School, University of Pennsylvania, 1967, and programed for the 360/50. The remaining comments are taken from the University of Chicago write-up of the program adapted by Marianne Stover: (1) The proportion of the total sum of squares that must be contained in any group if that group is to become a candidate for splitting is .0100 (this is the recommended value); (2) the best split on the *ith* candidate group must reduce the unexplained sum of squares by X proportion of the total sum of squares or that group will not be split, and it will not become a candidate group again even though it may meet the .0100 requirement above. The recommended value for samples of N = 200 or so is .01 for X (this is what we used). The other parameters involve the final group size and are discussed in the text.

[11] The random number program used in the study is called "Random" and is programed for the 7094. It is available from the Computation Center, University of Chicago.

TABLE 1

COMPARISON OF AID FOR DIFFERENT GROUP SIZES AS WELL AS FOR LINEAR REGRESSION

	AID (10 min.)				*AID (25 min.)*				*AID (50 min.)*				*Linear Regression*			
Sample Size	η	η^2	F	df	η	η^2	F	df	η	η^2	F	df	R	R^2	F	df
100	.54	.29	9.20*	5,93	.44	.19	12.83*	2,97	no split				.414	.172	1.84	10,89
204	.58	.34	8.23*	14,189	.41	.17	7.92*	6,197	.26	.07	9.02*	2,201	.274	.075	1.57	10,193
300	.54	.29	8.37*	17,282	.39	.15	8.29*	7,292	.32	.10	11.94*	3,296	.201	.041	1.22	10,289
404	.51	.26	9.78*	16,387	.35	.12	7.62*	8,395	.26	.07	8.07*	4,399	.184	.034	1.37	10,393
505	.30	.09	7.98*	7,497	.26	.07	7.44*	6,498	.24	.06	9.14*	4,500	.166	.028	1.40	10,494
605	.20	.04	6.85*	4,600	.17	.03	6.79*	3,601	.17	.03	6.79*	3,601	.152	.023	1.40	10,594
705	no split				no split				no split				.129	.017	1.17	10,694
805	no split				no split				no split				.111	.012	1.00	10,794
899	no split				no split				no split				.108	.012	1.05	10,888
999	no split				no split				no split				.106	.011	1.14	10,988

* $p < .001$. "No split" refers to the fact that the parent group was not split at all by the predictors.

$n = 800$, $F = 10.09$, $p < .001$). The data for the linear regression show that *none* of the results are significant. The data for group sizes of 25 and 50 indicate that increasing the group size does reduce the amount of explainable variance although there are still significant etas for sample sizes between 200 and 600 (the eta statistic is the correlation ratio and is used to measure nonlinear association between variables).

AID has been suggested as an exploratory device when the researcher has few hypotheses about possible relationships in the data. The results obtained here suggest that even for exploratory analyses, large samples are essential to guard against spurious splits and significant F values for random data. However, different researchers have different definitions of what constitutes a "large" sample. A conservative rule of thumb is to use at least 1,000 cases. If larger final group sizes are used, a smaller number of observations will suffice. These conclusions assume that there is *no* evidence from cross-validation. If the researcher has such evidence and finds that he can predict to new cases, the admonitions about sample size and final group size are superfluous.

ASPECTS OF MULTIPLE REGRESSION

One of the most widely used techniques in the social sciences is multiple regression. Several excellent papers are available that discuss regression from a particularly behavioral point of view.[12] This article discusses three sources of potential problems that may give rise to spurious inferences.

Consider the situation where the number of predictor variables is large relative to the size of the sample (because, for example, the researcher does not have any way of deciding which variables are really most important for predicting the criterion variable). Rather than drop a potentially valid predictor, he includes all the predictors in the regression equation. If the number of observations (N) available is not much larger than the number of predictors (k), the multiple R will be artificially high in the sense of being an overestimate of the population multiple correlation. (For $N = 2$, $k = 1$, R must be 1.0 since two data points determine a straight line; more generally, $R = 1.0$ if $k = n - 1$). If the question of interest is to obtain an estimate of the population multiple correlation, various "correction" formulas are available.[13] However, if one is interested in predicting to new

[12] J. Cohen, "Multiple Regression as a General Data-Analytic System," *Psychological Bulletin,* Vol. 69, 1968, pp. 426-443; R. B. Darlington, "Multiple Regression in Psychological Research and Practice," *Psychological Bulletin,* Vol. 69, 1968, pp. 161-182.

[13] H. Theil, *Principles of Econometrics,* New York, Wiley, 1971, pp. 178-179.

cases on the basis of the sample multiple regression equation, a large number of predictors relative to the sample size will result in large shrinkage. Darlington cites a case where 84 predictors were used in a study involving 136 subjects.[14] While initial fit yielded an $R = .73$, the cross-validated correlation was .04. A more dramatic example has been provided by Guion in the area of personnel selection:

> Consider, for example, the sad story of McCarty and Fitzpatrick (1956). Using the Wherry-Doolittle technique they selected a battery and estimated a shrunken $\bar{R}$ of .92. When they cross-validated on a second sample, however, they found the correlation to be $-.21$! . . . Considering the employment errors that would have been made had this battery been put to use for selection, one can easily see the necessity for cross-validation.[15]

It should be noted that, contrary to several sources, the Wherry shrinkage formula is inappropriate for estimating the cross-validated correlation.[16] At the moment, the only way to find out what this relationship will be is to do the cross-validation empirically.

Another source of potential problems in regression analysis concerns the use of "stepwise" regression. This technique uses the enormous facility of the computer to scan and select those predictors from a large set that will maximally correlate with the criterion of interest. The program works in the following manner. The computer selects the first predictor so that the correlation is the largest of any of the predictors with the criterion. Then variables are added, one at a time, on the basis of adding incrementally to the prediction of the criterion. This is done by computing *F*-tests for all the variables and choosing those that add most to the equation. While this technique may be quite useful in delimiting a subset of predictors from a large set, it is easy to include any and all variables and let the computer pick the "best" combination. The trouble is that the particular variables selected may be related to the criterion *in this sample only*. The variables that are chosen are capitalizing on chance within the particular sample. Standard significance tests and/or "adjustments" are not applicable for testing for statistical significance. If one is interested in predicting to a new set of cases, empirical cross-validation must be done. For example, Einhorn[17] found that for 10 predictors fit to 193 observations, the initial fit of a particular model was .430, which on cross-validation "shrank" to .380. When a stepwise procedure was used to pick out 10 predictors from 40 potential variables, the initial fit was

[14] Darlington, *op. cit.*, p. 174.
[15] R. M. Guion, *Personnel Testing*, New York, McGraw-Hill, 1965, p. 166.
[16] See Darlington, *op. cit.*, p. 173.
[17] H. J. Einhorn, "Expert Measurement and Mechanical Combination," *Organizational Behavior and Human Performance*, Vol. 7, 1972, pp. 86-106.

.549 but the cross-validated correlation was .363. Therefore, the cross-validated correlation was higher in the first case although the initial fit was higher in the second.

A new procedure, developed by Kruskal,[18] attempts to maximize the multiple correlation by programing the computer to perform various monotonic transformations on the predictors and the criterion. This procedure can be used without any theory as to what substantive meaning is to be attached to the transformations. While there has not been sufficient research on the ability of this program (and similar schemes that involve transformations on the variables in the regression analysis for purely statistical expedience), it is expected that such procedures will yield spurious relationships that will vanish on cross-validation. If one is interested in transformations for substantive reasons related to the particular problem and the general functional form is known, such procedures might prove useful. (Cf. Einhorn,[19] where a theoretical scheme was developed which implied certain transformations. These were carried out and held on cross-validation.)

FACTOR ANALYSIS

The literature on factor analysis is voluminous, and only certain aspects can be treated here (the interested reader is referred to Harman,[20] Eysenck,[21] Catell,[22] and Guilford[23]). The major goal of factor analysis (and the closely related procedure, principal components analysis) is to provide a set of factors—or constructs—that will account for the covariation among a set of variables. The usual input in the analysis is a correlation matrix (although raw scores can also be factored[24]). The output is a factor matrix, the rows being the variables of interest and the columns being the factors. The great advantage of the analysis is to obtain a small number of factors (smaller than the original number of variables) which will account for the linear dependencies (multicolinearities) between the variables.

Two papers of particular importance examine two different as-

[18] J. B. Kruskal, "Monotonic Multiple Regression." Program available at Computation Center, University of Chicago.

[19] H. J. Einhorn, "The Use of Nonlinear, Noncompensatory Models in Decision Making," *Psychological Bulletin,* Vol. 73, 1970, pp. 221-230.

[20] H. Harman, *Modern Factor Analysis,* 2d ed., Chicago, University of Chicago Press, 1967.

[21] H. J. Eysenck, "The Logical Basis of Factor Analysis," *American Psychologist,* Vol. 8, 1953, pp. 105-114.

[22] R. B. Catell, "The Three Basic Factor Analytic Designs—Their Interrelations and Derivatives," *Psychological Bulletin,* Vol. 49, 1952, pp. 499-520.

[23] J. P. Guilford, "When Not to Factor Analyze," *Psychological Bulletin,* Vol. 49, 1952, pp. 26-37.

[24] J. Nunnally, *Psychometric Theory,* New York, McGraw-Hill, 1967, ch. 11.

pects of factor analysis. The first, by Armstrong and Soelberg, questions the utility of factor analysis for dealing with real data, while the second, by Armstrong, attacks the idea that factor analysis can be used to *derive* theory.[25] The question first asked was: Do the applied studies using factor analysis lead to advances in the description or understanding of real situations? Armstrong and Soelberg provided the following "study": Fifty employees rated their supervisors on 20 traits. Pearson product-moment correlations were obtained between the ratings and the resulting matrix of correlations was factor-analyzed. Using standard criteria for extracting factors, they rotated the factors obtained for ease of interpretation. The analysis showed that 9 factors accounted for 71 percent of the variance contained in the correlation matrix. It was also quite easy to name the factors—i.e. the results seemed to make good conceptual sense. The only problem was that "employee responses" were *random normal deviates*! The arbitrary trait names applied to the factors were assigned *prior* to carrying out the analysis. The authors state:

> Since it appears to be rather simple for a researcher to make sense out of patterns provided by factor analysis, some benchmark or measure of reliability should be made a requirement for publication. Unfortunately, statistical tests for measuring factor reliability do not appear to be well developed.[26]

The authors go on to say that about two thirds of the studies they examined in various journals did not report a measure of reliability and 46 percent reported on neither reliability nor validity. (Reliability means that the factors would be obtained on a new set of responses, while validity means that the factors would correlate with some outside criterion.)

Armstrong and Soelberg discuss three potential ways to deal with assessing reliability: (1) split samples—this is similar to cross-validation and means that the analysis is done on each half of the sample to see whether the same results are obtained; (2) a priori analysis —this refers to having some theoretical scheme which can be compared to the obtained factors; and (3) Monte Carlo simulation—by factoring suitable samples of random data, sampling distributions can be obtained and used for comparing obtained results. This last method has received some attention in the literature.[27]

The second paper by Armstrong deals with a hypothetical analysis

[25] J. S. Armstrong and P. Soelberg, "On the Interpretation of Factor Analysis," *Psychological Bulletin,* Vol. 70, 1968, pp. 361-364; Armstrong, *op. cit.*

[26] Armstrong and Soelberg, *op. cit.*, p. 362.

[27] N. Cliff and C. D. Hamburger, "The Study of Sampling Errors in Factor Analysis by Means of Artifical Experiments," *Psychological Bulletin,* Vol. 68, 1967, pp. 430-445.

of a set of 63 metals by "Tom Swift", resident operations researcher for the American Metal Company. Although Tom knows nothing about metallurgy, he does know factor analysis. Swift has various measurements of the metals—thickness, density, weight, cost per pound, etc.—and he would like a typology of the metals based on their characteristics. He performs a factor analysis after correlating the metals over the various attributes. This factor analysis yields three factors that account for 90.7 percent of the variance, but the factors do not seem to be very meaningful. Another factor analysis is done with other variables, e.g. molecular weight, melting point, etc., added to the original variables. The results obtained were even more difficult to interpret. As Armstrong points out:

> The point is, however, that without a prespecified theory Swift has no way to evaluate his results . . . The factor analysis might have been useful in evaluating theory . . . Tom Swift's work would have been much more valuable if he had specified a conceptual model. He would have been able to present a more convincing argument for his resulting theory had it agreed with his prior model. Such agreement is evidence of construct validity.[28]

NONMETRIC MULTIDIMENSIONAL SCALING

The goal of nonmetric multidimensional scaling is to find a spatial configuration for n stimuli such that the rank order of the distance between the n stimuli is maximally inversely related to the rank order of the proximity measures (similarity judgments) obtained empirically. The scaling is called nonmetric because only the rank-order information is necessary to perform the scaling. In order to do a scaling analysis, it is necessary to obtain similarity or dissimilarity judgments for all possible pairs of the n stimuli, i.e., $n(n - 1)/2$ pairs. The scaling differs from univariate scaling in that the stimuli being scaled are quite complex—as a matter of fact, the chief concern of multidimensional scaling is to determine *how many dimensions* the subjects are using in making their judgments of similarity.

Although there are a number of scaling procedures, only Kruskal's version will be discussed.[29] However, most of the available methods are quite similiar. The procedure starts with an arbitrary configuration of points and iteratively attempts to find a spatial configuration such that the rank order of the interpoint distances shows a perfect negative correlation with the rank order of the similarity measures. The procedure attempts to do this while reducing the dimensionality of the space. As the dimensionality gets smaller, the ability of the program

[28] Armstrong, *op. cit.*, p. 20.

[29] J. B. Kruskal, "Multidimensional Scaling by Optimizing Goodness of Fit to a Nonmetric Hypothesis," *Psychometrika*, Vol. 29, 1964, pp. 1–27.

to get perfect relationships decreases. Kruskal proposed a measure of departure from perfect fit called "stress". Stress can be conceived of broadly as residual variance or analogous to the standard error of estimate in regression analysis.

In order to find the best number of dimensions for the analysis, the stress measure is computed for the various dimensionalities. Stress can be plotted as a function of the number of dimensions and the researcher must decide what the proper number of dimensions is for his data. Although Kruskal gives three criteria for making this judgment, these criteria cannot always be applied.[30]

The most important aspect of the scaling procedure, from our perspective, is the statistical significance of the result. This was the question asked by Klahr,[31] who performed a Monte Carlo simulation of the scaling algorithm on random data for $n = 6, 7, 8, 10, 12$ and 16. This means that the number of pairs of judgments was 15, 21, 28, 45, 66, and 120 respectively. For $n = 6, 7, 8$, and 10, one hundred sets of random numbers were generated while for $n = 12$ and 16, 50 sets were used. Every set was scaled using Kruskal's method, results yielding spaces (dimensionalities) of from 1 to 5 dimensions. It should be noted that while an n of 6 is small, this still requires 15 judgments if the analysis is done with actual subjects. Therefore, the values of n used in this study should be highly representative of the range of most studies that are feasible with human subjects.

The results of the analysis showed that stress was highly sensitive to n. "Good" solutions (i.e., stress $\leq .05$) were often available for 6, 7, and 8 points in three dimensions when the proximity measures were *randomly generated*. However, an increase to $n = 10$ greatly reduced the likelihood of this kind of spurious result. ($N = 10$ means that 45 judgments of similarities would have to be made—not an altogether easy task for complex stimuli. It may also be that the unreliability of the judgments increases as the number of judgments increases.) The basic findings of the study are summarized by Klahr:

> The importance of these findings to a user of Kruskal's program rests upon the manner in which the scaling procedure is being used. If it is being used to test *a priori* hypotheses about the dimensionality or spatial arrangement of a stimulus set, then the significance criterion is but one of many pieces of evidence that can be used in interpreting results. However, if the scaling procedure is used in an exploratory study, where there are no a priori notions about the configurations, then any result for only 8 or 9 points in 2 or 3

[30] H. H. Stenson and R. L. Knoll, "Goodness of Fit for Random Rankings in Kruskal's Nonmetric Scaling Procedure," *Psychological Bulletin,* Vol. 71, 1969, pp. 122-126.

[31] D. Klahr, "A Monte Carlo Investigation of the Statistical Significance of Kruskal's Nonmetric Scaling Procedure," *Psychometrika,* Vol. 34, 1969, pp. 319-330.

dimensions cannot be considered very convincing evidence for the existence of structure. For example, from Figure 3 it is evident that with eight points there are about 2 chances in 3 that pure noise could be accounted for in 3 dimensions with a stress less than .075.[32]

CONCLUSIONS

This article has discussed the potential of several analytic techniques for generating spurious results. It should be clear that proceeding *without* a theory and *with* powerful data analytic techniques can lead to large numbers of Type I errors. Just as the ancient alchemists were not successful in turning base metal into gold, the modern researcher cannot rely on the "computer" to turn his data into meaningful and valuable scientific information.

[32] *Ibid.*, pp. 328-330.