Golden, R., Tyan, S., & Meehl, P. E. (1974). *Detecting latent clinical taxa, IX: A Monte Carlo method for testing taxometric theories* (Report No. PR-74-7). Minneapolis: University of Minnesota, Research Laboratories of the Department of Psychiatry.

Reports from the Research Laboratories

of the

Department of Psychiatry

University of Minnesota

Detecting Latent Clinical Taxa, IX:

A Monte Carlo Method for

Testing Taxometric Theories[1]

Robert R. Golden, Shirley H. Tyan

and

Paul E. Meehl

Report No. PR-74-7 October 1974

[1] This Research was supported in part by grants from the Psychiatry Research Fund and the National Institute of Mental Health, Grant Number MH 24224

TABLE OF CONTENTS

[This file contains the text only. Appendices are in file 104TechRep9Appendices.pdf]

I. Basic Background Rationale

The parameter estimation procedures of the taxometric theories are complemented by consistency tests which describe how well the theory fits the data. The major purposes of consistency tests are to avoid spurious taxometric findings (e.g., the method indicates there is a taxon when, if fact, there is nothing of the sort) and to detect when parameter estimates are too erroneous for the particular purposes of the study. Consistency testing in application of a mathematical theory is just as obviously required and is just as much a matter of simple common sense, as in other endeavors. For example, the builders of the MMPI realized that validity keys were required. However, while anyone would know that some people randomly respond and lie and so on when taking the MMPI, it is curious that few psychologists or sociologists act as if Nature could, on occasion, be more devious than mathematicians require.

In general, any taxometric theory can be thought of as a set of equations relating a set of latent parameters to a set of manifest parameters. Some of these equations may involve only latent parameters and others involve only manifest parameters. Most equations of most immediate concern in the development of a theory involve both kinds of parameters.

There are two special types of equations: (A) the assumptions and (B) the derived equations which express the latent parameters as explicit functions of the manifest parameters. Traditionally, the psychometrician usually is satisfied with just the development of B from A. while such a feat may require a high degree of mathematical

competence and creativity and can be regarded as the solutions of the most immediate importance, there remains further mathematical derivation to prepare the theory for application to substantive problems. Such derivation can be roughly described to be that of deriving all further relations between the parameters that one is able to. These latter equations can be used for determining how well the theory fits the data of the real phenomena, hence are called the (C) consistency equations. If the assumptions A are roughly correct and the estimates of the manifest parameters (obtained from the data, of course) and of the latent parameters (by the calculations given by B) are roughly correct, then substitution of the parameter estimates into C will show that they roughly satisfy each equation of type C.

There are at least four sources of error which cause the consistency equations to be only approximately satisfied. First, the assumptions A are always mathematical idealizations and never strictly true for real phenomena, and therefore it is clear that the estimates resulting from B will contain such error. Second, the manifest parameters contain sampling error (between individuals) and measurement error (within individuals); hence, the latent parameter estimates contain sampling and measurement error (since they are functions of the manifest parameters as given by B). Third, the calculation method used in B can be one that according to the underlying mathematical theory gives at best an approximate solution to the equations resulting from A.

Theoretically, sampling error, measurement error and solution error can be assessed rather directly and can be reduce to (nearly) any arbitrarily small size. To reduce

(a) sampling error, one can increase the sample size, (b) measurement error, one can resort to reliability theory, factor analysis, and item selection methods or (c) solution error, one can, for example, continue an iterative calculation procedure until convergence conditions are adequately satisfied. However, "assumption error" does not seem to be of this same sort in that its size cannot be *directly* assessed (an assumption as opposed to a hypothesis is by definition not directly testable) and it is not reducible to (nearly) any arbitrarily small size by a systematic procedure. Naturally there is no corresponding theory of verisimilitude to numerically assess assumption error.

Suppose that only assumption error is a matter of concern; that is, all other sources of error have been eliminated. Presumably continual revision of the theory so that the parameter estimates of B become closer to perfect solutions of C would increase the verisimilitude of the theory assuming that the consistency equations are chosen correctly so as to provide sufficient testing of the fit of the theory to the data. The consistency testing development might then attempt to meet criteria such as the following:

(a) there is one for as many subsets of the assumptions as possible,

(b) they are not redundant in that they are derivable from B; even the addition of weak assumptions to B should not allow the derivability of C,

(c) they follow as directly from A as does B and, in fact, might be partially interchangeable with B,

(d) they are adequately sensitive to assumption errors that are most probable,

(e) they are adequately sensitive to assumption errors that are most troublesome in that they cause intolerable errors in important parameter estimates,

(f) they provide clues as to how the theory might be revised to obtain a better fit (by pointing out the set of disparate assumptions with the aid of criterion (a)), and

(g) they indicate when the theory is totally off the mark and should not be used at all.

With the current state of the art of mathematical theory building in the area of psychopathology measurement it would be a major contribution to meet even the last of these criteria.

Unfortunately, formal mathematical analysis in the way of determining whether or not consistency tests satisfy the above criteria for a complicated taxometric theory is always terribly difficult and usually impossible in a practical sense. Fortunately, however, the Monte Carlo method can be used to perform an approximative analysis which is good enough for nearly all purposes. In this method, unknown mathematical functions are approximately described by observing the behavior of the dependent variables resulting from systematic variation in the independent variables. Most of the above criteria for evaluating consistency tests can be given in terms of properties of various mathematical functions. Even though these functions are terribly complicated when given explicitly or implicitly, they can be easily studied for our purposes by the Monte Carlo method.

II. General Design of the Monte Carlo Experiment

The basic idea of the design proposed here is very simple. Suppose θ_i are the (population) parameters of a particular taxonomic structure. Suppose further, that we are able to construct an artificial data (variable $\times$ subject) matrix which can be regarded as a sample from some specified population given in terms of θ_i's. We can then analyze the artificial data by the taxometric method and observe the errors (ε_i)[1] of the parameter estimates ($\hat{\theta}_i$) and the degree of approximations (Δ_j) obtained in the consistency equations. It will be noted that for the taxometric theory to work correctly it is necessary that:

(a) the ε_i's and the Δ_j's be sufficiently small (the ε_i's are set in accordance with the substantive problem estimation accuracy requirements and the Δ_j by a method given below) or

(b) if not, one or more of the Δ_j's exceed critical values c_j (one or more consistency tests are failed).

The critical values (c_j) are usually determined analytically in the following manner. The consistency equation can usually be written in the form

$$F_j (\theta_1, \theta_2, \ldots \theta_n) = 0.$$

Then the exact differential is given by

$$dF_j = \sum_i \frac{\partial F_j}{\partial p_i} d\theta_i \ ;$$

[1] Let a_i denote an arbitrary but fixed vector *element*, then the notation (a_i) will be used to denote the vector $(a_1, a_2, a_3, \ldots a_n)$.

hence if we define Δ_j by

$$\Delta_j = \sum_i \frac{\Delta F_j}{\Delta p_i} \varepsilon_i$$

we can determine what Δ_j will be for small θ_i errors ε_i for any given value of θ_i. Usually it is sufficient to choose for the cut-value c_j an upper bound for the Δ_j values which result from considering maximally tolerable values for the $\Delta\theta_i$ and a sufficiently wide variety of θ_i values; this procedure normally can be handled analytically but the Monte Carlo method can be resorted to if necessary.

To study sampling error for a particular value of θ_i a number of independent data samples are generated and analyzed and appropriate $\varepsilon_i \times \Delta_j$ scatter plots are studied to see if conditions (a) and (b), given above, are met.

One is not interested in one value of θ_i but in all values which are at least remotely possible. The sample size is made sufficiently large so that it need not be a matter of concern. Here each point of a $\varepsilon_i \times \Delta_j$ scatter plot represents a value of θ_i. The problem then is to select a finite and reasonably small set of θ_i values which give a sufficiently accurate picture of the $\varepsilon_i \times \Delta_j$ relationship for all reasonably likely taxonomic situations (or θ_i values). There is no firm analytical procedure offered here for adequately sampling taxonomic situations but as will be discussed below, the method does not ultimately rely on an exhaustive sampling.

It was implied above that the $\varepsilon_i \times \Delta_j$ relationship be viewed in terms of a four-fold table. However, to the extent that one's substantive theory testing is a matter of determining the degree of fit within the context of discovery rather than making

a dichotomous decision (accept, reject) it would be better to record the entire $\varepsilon_i \times \Delta_j$ scatter plot or summary statistics of it.

Each sample of a taxonomic situation selected produces a value of ε_i and a value of Δ_j. An interesting possibility for determining major causes of error in a taxometric method consists of factor analysis of ε_i and Δ_j variables, the ε_i variables alone and the Δ_j variables alone. In this instance we will *know* if the factors are real, interpretable, and useful for if they are they will lead to new insights about the method.

III. The Requirement of the Random Number Generator Method

It remains to determine how the artificial data are to be generated. In this section are discussed the various primary considerations which lead to the development of the proposed generating method.

The various latent taxonomic detection methods developed so far use either scales (for example, see Golden *et al*., 1974b) or dichotomous items (for example, see Golden *et al*., 1974c) as indicators. In practice, usually the scales will simply be sums of dichotomous items such as those of the MMPI (called keys herein). The methods usually rest on an assumption which requires that correlations between the indicators be zero within each taxonic class. In the case of keys this condition is met if the items of one key are each independent of those of another. The normal theory (Golden *et al*., 1974b) — both the maximum likelihood and minimal chi-square solutions — requires that the within-taxonomic class distributions be normal; this condition also is obtained when the

items of the key are independent within taxonomic class. In general, all of the major assumptions of any of these theories are derivable from the grand assumption that the within-taxonomic class item correlations are all perfectly zero.

In reality, items are never perfectly independent within taxonomic class, of course, and it is necessary to know how robust a method is with respect to this assumption. That is, it is desirable to know how highly and in what ways or patterns items can be correlated within taxonomic class without seriously affecting the method's estimates of the latent parameters describing the taxonomic situation.

The problem then is to generate artificial data that simulate *real* Bernoulli variables with specified degrees and patterns of dependency. Since the assumptions of the various taxometric theories can all be given in terms of phi-correlations between items within taxonomic class, it should be simple to specify these. Thus, the simple-to-use method proposed here requires only the specification of the single parameter for each Bernoulli variable (the mean or proportion) and the correlation matrix for the set of Bernoulli variables.

IV. Analytical Development of a Random Number Generator Method

Suppose that $\tilde{z}_i$ are n Bernoulli random variables which have the value 1 with probability p_i[2] and the value 0 with probability $1 - p_i$, i = 1, 2, 3, …, n. Let the joint proportion matrix be denoted by

[2] In this development *all* constants are population parameters.

$$\underset{n\times n}{p} = \begin{bmatrix} p_1 & & & \\ p_{21} & p_2 & & \\ p_{31} & p_{32} & p_3 & \\ \vdots & & & \\ p_{n1} & p_{n2} & \cdots\cdots & p_n \end{bmatrix}$$

the phi-correlation coefficient matrix by $\underset{n\times n}{\phi} = \phi_{ij}$, where

$$\phi_{ij} = \frac{p_{ij} - p_i p_j}{\sqrt{p_i\left(1-p_i\right)p_j\left(1-p_j\right)}}$$

and the vector of variable parameters $\underset{1\times n}{p} = (p_1, p_2, \ldots, p_n)$. The problem, then, is to generate numbers simulating Bernoulli variables with given $\underset{1\times n}{\phi}$ and $\underset{1\times n}{p}$ matrices, or, equivantly, with a given $\underset{n\times n}{p}$ matrix.

Let $\tilde{x}_i$ (i = 1, 2, 3, …, n) be n standardized normal random variables (with mean equal to zero and unit variance) with intercorrelations ρ_{ij} (where i,j = 1, 2, 3, …, n). Then we construct n Bernoulli variables $\tilde{z}_i$ by

$$\tilde{z}_i = \begin{cases} 1 & \text{if } F\left(x_i\right) \le p_i \\ 0 & \text{if } F\left(x_i\right) > p_i \end{cases} \qquad (1)$$

where

$$F\left(\tilde{x}_i\right) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x_i} e^{-t^2/2}\, dt \ .$$

It follows that

$$\rho_{ij} = E\left(\tilde{z}_i\ \tilde{z}_j\right) = \Pr\left(\tilde{z}_i = 1 \text{ and } \tilde{z}_j = 1\right)$$

$$= \Pr\left(F\left(\tilde{x}_i\right) \le p_i \text{ and } F\left(\tilde{x}_j\right) \le p_j\right)$$

$$= \frac{1}{2\pi\left(1-\rho_{ij}^2\right)^{1/2}} \int_{-\infty}^{p_i}\int_{-\infty}^{p_j} \exp\left\{-\frac{1}{2\left(1-\rho_{ij}\right)}\left(x^2 - 2\rho_{ij}xy + y^2\right)\right\} dxdy \qquad (2)$$

which is the bivariate normal density function for two standard variables (p. 33, Kendall 1952).

Since

$$E\left(\tilde{z}_i\ \tilde{z}_j\right) = p_{ij} = \phi_{ij}\left\{p_i\left(1-p_i\right)p_j\left(1-p_j\right)\right\}^{1/2} + p_i p_j\,, \qquad (3)$$

it follows that given ϕ_{ij}, p_i and p_j, which are the parameters we wish to specify, then p_{ij} can be determined by (3) and, next, ρ_{ij} is determined by solution of (2) for ρ_{ij}. Suppose it were possible to solve (2) explicitly for ρ_{ij} such that

$$\rho_{ij} = F(\phi_{ij}, p_i, p_j),$$

then given ϕ_{ij} and $(p_1, p_2, \ldots, p_n)$ one could determine ρ_{ij}. Leaving this matter of determining ρ_{ij} for the moment, we observe that since the $\tilde{x}_i$ are distributed multivariate normally they can be generated by a method such as the one presented below. The cumulative density for each value of each $\tilde{x}_i$ is then determined and is compared with p_i according to (1) to determine the value of the Bernoulli variable $\tilde{z}_i$. This completes the process of generating the $\tilde{z}_i$ values for specified $\underset{n\times n}{\phi}$ and $\underset{1\times n}{p}$ matrices.

It remains, first, to describe a method for simulation of a multivariate normal distribution given the correlation matrix and the variable means and variances.

Let $\underset{n\times 1}{\tilde{x}} = (x_1, x_2, \ldots x_n)$ be an n-tuple random vector such that each $\tilde{x}_i$ is a standardized random variable (with zero mean and unit variance) and let the covariance between $\tilde{x}_i$ and $\tilde{x}_j$ be σ_{ij} for each i,j pair and define

$$\underset{n\times n}{\sum} = \begin{bmatrix} 1 & \sigma_{12} & \cdots & \sigma_{1n} \\ \sigma_{21} & \ddots & & \\ \vdots & & \ddots & \\ \sigma_{n1} & & & 1 \end{bmatrix}$$

as the given covariance matrix. Let $\underset{n\times 1}{y}$ be a set of such variates such that the covariance matrix is I, the identity matrix. Suppose that there exists a matrix $\underset{n\times n}{T}$, for a given Σ, such that $\underset{n\times 1}{x} = \underset{n\times n}{T} \cdot \underset{1\times n}{y}$. Thus

$$\underset{n\times n}{\Sigma} = E\left(\underset{n\times 1}{x} \cdot \underset{1\times n}{x'}\right) = E\left\{(T\cdot y)\cdot(Ty)'\right\}$$

$$= E(Ty \cdot y'T) = TE(yy')T' ;$$

since $E(yy') = I$, we have $\underset{n\times n}{\Sigma} = \underset{n\times n}{T} \cdot \underset{n\times n}{T'}$

and it is seen that T is the matrix such that the product of it and its transpose is the given Σ. When an unique solution for T exists it can be found by a method given by Anderson (1957). The y_i can be simulated by use of any of several well-known methods for univariate normal variables.

Finally, it remains to determine ρ_{ij} since it is not easily found as an explicit function of ϕ_{ij}, p_i and p_j. The function is adequately approximated by tables of corresponding values of p_{ij}, ϕ_{ij} and ρ_{ij} (given in Appendix B). These tables were constructed by specifying values of p_i, p_j and ρ_{ij} and determining p_{ij} and ϕ_{ij} from (2) and (3).

The parameters p_i and p_j were allowed to vary from 0.1 to 0.9 in increments of 0.1 and p_{ij} from 0 to min(p_i, p_j) in increments of 0.025. The integration in (2) was performed by a numerical method (see pp. 887 and 916 of Abramowitz and Stegan, 1964).

This completes the development of the method. The basic idea is simply to use the multivariate normal distribution, dichotomize by cutting deviates corresponding to cummulative densities p_i and to preadjust the ρ_{ij} values for the resulting shrinkage caused by dichotomization.

In order to use the generator it is necessary to specify ϕ_{ij} and $\underset{1\times n}{p}$ then use the table to determine ρ_{ij} which is required input for the subroutine (described in Appendix A). It should be noted that not all values for ϕ_{ij} and $\underset{1\times n}{p}$ are acceptable. A straightforward procedure is to pick the p_i values first, then pick ϕ_{ij} values such that

$$p_i p_j \leq p_{ij} \leq \min(p_i, p_j)$$

or

$$0 \leq \phi_{ij} \leq \frac{\min\left(p_{ij}p_j\right) - p_i p_j}{p_i\left(1-p_i\right)p_j\left(1-p_j\right)}.$$

It can also be shown from a result given by McNemar (p. 166, 1962) that it is required that for each i,j pair

$$\sqrt{1-\phi_{ik}^2}\sqrt{1-\phi_{jk}^2} - \phi_{ik}\phi_{jk} < \sqrt{1-\phi_{ik}^2}\sqrt{1-\phi_{jk}^2} + \phi_{ik}\phi_{jk}$$

for all k ($k \neq i \neq j$). Although the above two conditions are necessary ones, it has not been shown that they are sufficient.

Example

Suppose we wish to generate numbers for five Bernoulli variables with the following parameter values:

$$p_i = (.320, .780, .880, .100, .510)$$

and

$$\phi_{ij} = \begin{matrix} 1.00 & & & & \\ 0.32 & 1.00 & & & \\ 0.31 & 0.37 & 1.00 & & \\ 0.29 & 0.15 & 0.12 & 1.00 & \\ 0.34 & 0.09 & 0.26 & 0.06 & 1.00 \end{matrix}$$

From appendix B we find that

$$\rho_{ij} = \begin{matrix} 1.00 & & & & \\ 0.68 & 1.00 & & & \\ 0.87 & 0.64 & 1.00 & & \\ 0.57 & 0.49 & 0.65 & 1.00 & \\ 0.53 & 0.17 & 0.54 & 0.13 & 1.00 \end{matrix}$$

The observed sample correlation matrices for two independently generated samples of size 1000 are given below.

Sample 1

$$\hat{\phi}_{ij} = \begin{matrix} 1.00 & & & & \\ 0.34 & 1.00 & & & \\ 0.26 & 0.40 & 1.00 & & \\ 0.31 & 0.18 & 0.13 & 1.00 & \\ 0.33 & 0.10 & 0.23 & 0.05 & 1.00 \end{matrix}$$

Sample 2

$$\hat{\phi}_{ij} = \begin{matrix} 1.00 & & & & \\ 0.27 & 1.00 & & & \\ 0.25 & 0.32 & 1.00 & & \\ 0.31 & 0.16 & 0.13 & 1.00 & \\ 0.33 & 0.10 & 0.27 & 0.07 & 1.00 \end{matrix}$$

Comparison of the sample matrices with specified population illustrates the fact that the method works accurately enough for study of taxometrlc methods to be used in psychopathology. It is true, however, that some of the elements of these matrices depart significantly from the specified values. This is apparently due to error in the integration to obtain ϕ_{ij} as a function of ρ_{ij} and error from the interpolation required to obtain ρ_{ij} values for corresponding ϕ_{ij} values.

The method is essentially the same as the multivariate normal one except that each standard normal variate y_i is dichotomized by a cut at p_i. The only problem results from the fact that any given value for the ρ_{ij} matrix does not result in the same value for the ϕ_{ij} matrix (except in the special case where they each are the identity matrix) as the latter matrix elements are very roughly two-thirds the size of the former ones. Thus, it is simply necessary to adjust the ρ_{ij} values for this shrinkage so that the ϕ_{ij} values are of the size desired.

When p_i and p_j are both between .30 and.70 and ϕ_{ij} is between 0 and 0.5, there appears to be an approximate linear relation between ρ_{ij} and ϕ_{ij} given roughly by $\phi_{ij} = .60\rho_{ij}$ or $\rho_{ij} = 1.67\phi_{ij}$ although it has not been determined how accurate the relationship is. It would be interesting to sample uniformly from the lines of the table in Appendix B and determine the accuracy of the multiple regression equation for predicting ρ_{ij} from ϕ_{ij}, p_i and p_j. If everywhere accurate enough such an equation could replace the table. Possibly a more promising way to express ρ_{ij} as a function of ϕ_{ij}, p_i and p_j,

however, is to use the formula for the tetrachloric correlation coefficient which can be put in the form $\phi_{ij} = f(\rho_{ij}, p_i, p_j)$ where f is an infinite series. To obtain f^{-1}, Newton's or Horner's methods can be used according to McNemar (pp. 193ff, 1962).

V. Specification of the Taxonomic Class Distribution

There are at least three major properties of a taxometric theory which are of interest to study. First, it is of interest to determine the power of the method for accurately detecting taxonomies which have (a) various degrees of taxonomic class distribution overlap and (b) various degrees and kinds of assumption departures. Second, it is of interest to determine the ability of the method to avoid spurious taxon detection under any conditions. Third, it is of interest to determine if the consistency tests correctly identify accurate vs. inaccurate parameter estimates, and spurious vs. non-spurious detection. According to the nature of the derivation of the formulae for the estimation of the latent parameters, it would appear to clearly follow that if the assumptions are satisfied well enough then the method will work well enough with respect to the above three matters of concern. In other words, to study the major properties of a taxometric theory requires only a study in what happens under systematic departure from the ideal conditions specified by the assumptions. Hence we next consider how the p_i and ϕ_{ij} values can be selected in order to do this.

As the generator method is intended to be used separately for each taxomomic class, the various properties of the indicator distributions will be considered as functions of the specifiable parameters p_i and ϕ_{ij} for just a single taxonomic class.

a) Indicator Mean

The mean of a key is simply Σp_i.

b) Indicator Variance

The variance of a key is given by

$$\sigma^2 = \Sigma p_i q_i + 2 \sum_{\substack{i,j \\ (i \neq j)}} c_{ij}$$

where c_{ij} is the covariance between items i and j. This formula can be rewritten as

$$\sigma^2 = \Sigma p_i q_i + 2 \sum_{\substack{i,j \\ (i \neq j)}} \phi_{ij} \sqrt{p_i (1-p_i) p_j (1-p_j)}$$

If $p_i = p$ for all i then

$$\sigma^2 = npq + n(n-1)pq\bar{\phi}$$

where $\bar{\phi}$ is the average of the ϕ_{ij}'s. It is seen, then, that when the variance of a key is to be controlled it is helpful to have either $\phi_{ij} = 0$ or $p_i = p$.

By systematic variation of the means and variance of each taxonomic class distribution, the effects of taxonomic class distribution overlap and ratio of taxonomic class variances can be studied.

c) Indicator Skewness and Kurtosis

If the ϕ_{ij} are all zero then the resulting distribution is quasi-normal. It is not necessary that the p_i's be equal (for proof see p. 24ff, Golden and Meehl, 1973a). As the ϕ_{ij} are made to depart positively from zero in various ways, evidently, the distribution can be made skewed in either direction, platykurtic or leptokurtic, bimodal, U-shaped or J-shaped.

d) Correlation Between Two Indicators

If the Indicators x and y are keys of length n, with average within-key interitem correlations $\bar{\phi}_{wx}$ and $\bar{\phi}_{wy}$, and with an average between-key interitem correlation $\bar{\phi}_{bxy}$ then McNemar (p. 207, 1962) showed that the correlation between the two keys is given by

$$r_{xy} = \frac{n^2 \bar{\phi}_{bxy}}{\sqrt{n + (n-1)\bar{\phi}_{wx}}\ \sqrt{n + (n-1)\bar{\phi}_{wy}}}$$

when all items of both keys have the same p_i value. Thus, by controlling $\bar{\phi}_{wx}$, $\bar{\phi}_{wy}$ and $\bar{\phi}_{bxy}$ we can control r_{xy}. In Figure 1 where $\bar{\phi}_{wx} = \bar{\phi}_{wy} = \bar{\phi}_{bxy} = \bar{\phi}$ it is seen, for example, that for 20 item keys it is necessary to keep $\bar{\phi}$ below .05 in order that r_{xy} be less than 0.5 which is usually required by the maximum covariance theory (Golden and Meehl, 1973b). According to limited Monte Carlo evidence, the normal theory (Golden et al., 1974b) requires that r_{xy} be less than 0.8 or $\bar{\phi}$ be less than .20.

The base-rate of a taxonomic class is handled differently than the other parameters. Suppose that the compound sample size is to be N and the population taxonomic class proportion is p_j. The sample proportion $\hat{p}_j$ is distributed normally with mean p_j and variance $p_j(1 - p_j)/N$; hence a univariate normal random number generator can be used to obtain $\hat{p}_j$ and $N_j = \hat{p}_j N$.

In specifying the p_i's and the ϕ_{ij}'s there is a more subtle consideration which must be made. This consideration results from the fact that the number of times the generator is used to create a data sample to be mixed with the others is not necessarily the number of taxometric classes if the taxonomic situation is as described by the taxometric theories mentioned herein. It could be that p_i and ϕ_{ij} values chosen are approximately those

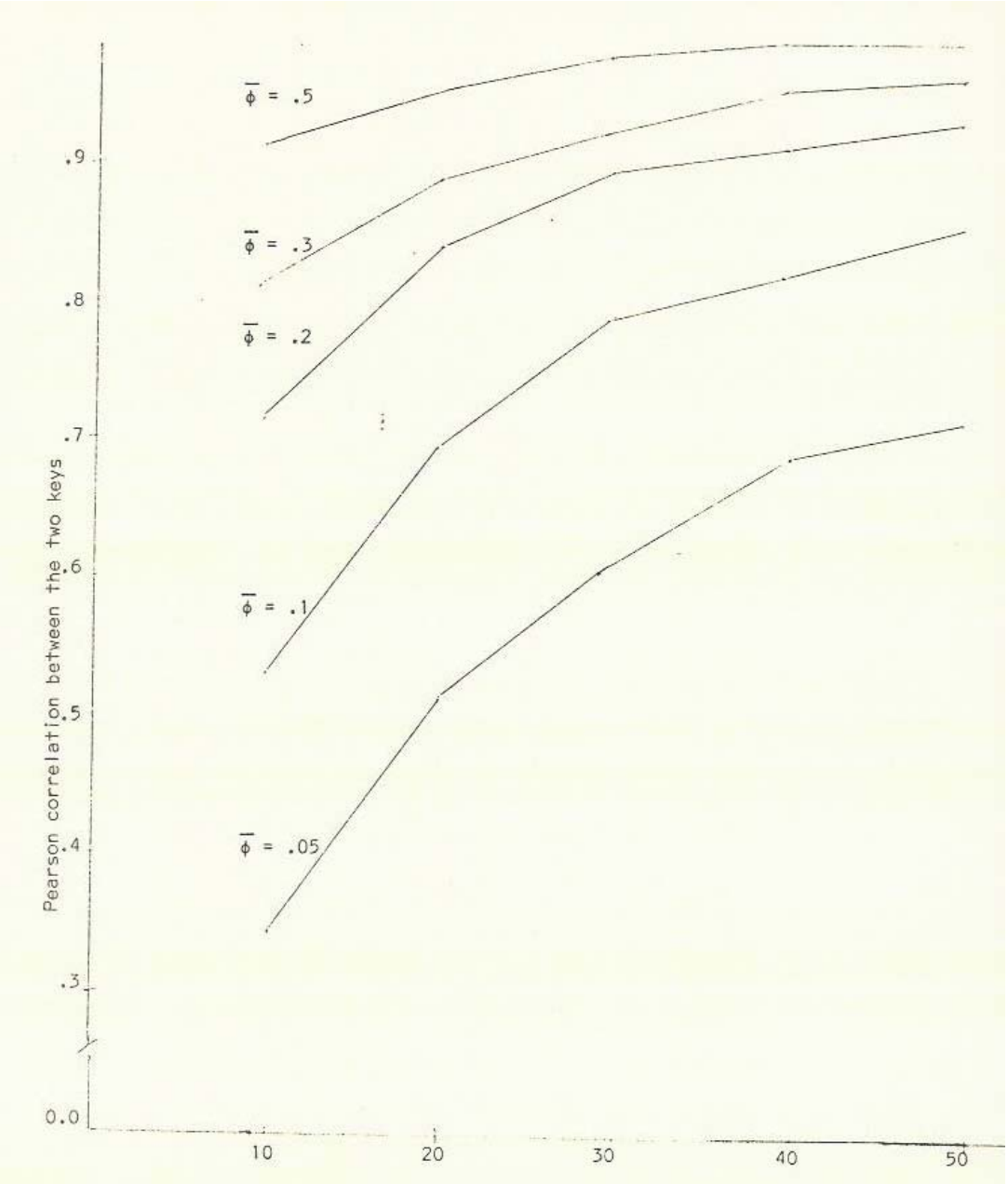

Figure 1. The Pearson correlation between two keys as a function of the common keylength for different values of the common average interitem correlation within key, $\bar{\phi}$.

of, say, a compound group consisting of two (or more) taxonomic classes. If such an error is made, then, among other things, the actual number of taxonomic classes is at least one more than it is thought to be. To demonstrate this possibility, suppose that a compound group consists of two taxonomic classes, denoted by subscripts 1 and 2, one with base-rate P, and within each class the item covariances are all zero. Then the covariances for the compound group are given by

$$c_{ij} = PQ(p_{i2} - p_{i1})(p_{j2} - p_{j1}).$$

Thus, if there exist values of P, p_{i1}, and p_{i2} (i = 1, 2, …, n) which satisfy the equations

$$\phi_{ij}\, p_i(1 - p_i)\, p_j(1 - p_j) = PQ(p_{i2} - p_{i1})(p_{j2} - p_{j1}),$$

$$p_i = P\, p_{i2} + Q\, p_{i1} \quad \text{and}$$

$$p_j = P\, p_{j2} + Q\, p_{j1}$$

where the p_i's, p_j's and ϕ_{ij} 's are specified, then the data of two taxonomic classes are simulated rather than that of one. It should be noted that the equations need be satisfied only in the approximate sense as would be the case for two real taxometric classes. There are $\binom{n}{2} + n = \frac{n(n+1)}{2}$ such equations whereas the number of unknowns P, p_{i2}, and p_{i1} is only 2n + 1. Therefore, the system is usually well overdetermined and a solution should not accidentally happen to exist due to sampling error. If a method does produce such a too-many-taxa result, the above equations can be resorted to in order to see if the method-produced suspected parameter estimates are approximative solutions.

How one can best take into account all the relations concerning the ϕ_{ij}'s and the p_i's simultaneously remains to be determined. Usually there is no problem in picking the p_i's, but the selection of the ϕ_{ij}'s can become complex. We typically want to control simultaneously for variances, general distribution shapes, covariation between indicators, and covariation within indicators as each has to do with a kind of manifestation of departure from the ideal condition in which all the ϕ_{ij}'s are zero. It is necessary that indicator distributions be specified in terms of skewness, variance, intercorrelatlon, homogeneity and the like as these are the parameters, in terms of which we have some knowledge as to the nature of real distributions. Further, we cannot be concerned about how well the taxometric method works for every single set of values of the ϕ_{ij} matrices even with each matrix element restricted to just three values as there are just far too many combinations to make this feasible. There are two queries then: "What are the algebraic procedures specifying the ϕ_{ij}'s for getting any desired multi-indicator distribution given in terms of skewness, homogeniety, and the like?" and "How can such procedures be made quick and simple to use?". One approach would be to select ϕ_{ij} and p_i values semi-randomly and then record the resulting indicator distributions and descriptive statistics along with the p_i and ϕ_{ij} values. (We assume here again that N is sufficiently large so that sampling error need not be of concern.) This procedure would be continued until either we have a catalogue of every kind of multi-indicator distribution we might wish to use or until we are not able to generate a new kind of distribution. This may seem like an ambitious task but it would appear that there are not more than a few hundred qualitatively different appearing distributions for two or three indicators.

VI. Simulation of Real Data

Another question about the method stems from the obvious fact that it does not allow or require use of all of the degrees of freedom available in specifying a taxonomic class distribution. A population taxonomic class distribution is uniquely specified in terms of the proportions of individuals for each of the 2^n different vectors of item responses (z_1, z_2, z_3, …, z_n). Since items are restricted to be dichotomous, we have $2^n - 1$ degrees of freedom in specifying a distribution. The random number generator method requires specification of only $\binom{n}{2}$ independent values of ϕ_{ij} and n independent values of p_i or $\frac{n(n-1)}{2} + n = \frac{n(n+1)}{2}$ independent parameter values all together. The quantity $2^n - \frac{n(n+1)}{2} - 1$ is greater than zero for $n \geq 3$ and is very large for common values of n in taxometric work. For example, if n = 10, the quantity is 1024 – 55 – 1 = 968 and there are 968 degrees of freedom used in some unknown manner by the generator method. In other words, the method imposes (hidden) constraints of an unknown nature that require a large number of degrees of freedom. This leads one to wonder just what sort of distributions can be generated by the method. This question remains even though the discussion of the preceding section indicates that parameters can somehow be specified to obtain any sort of assumption departure one might conceivably desire. The reason being that the familiar properties of distributions such as variance, kurtosis, homogeneity, and

the like may not include all of the important ones to be considered with regard to assumption violation for a taxometric theory.

Since what we ultimately wish to do is simulate real distributions, it would be of interest to see if this can be done for a large variety of real taxa, real non-taxa and, since there is a shortage of proven instances of these, of real criterion group distributions. The comparison would be in terms of (a) the familiar descriptive statistics and (b) the various important taxonomic parameter estimates resulting from analyzing compound groups of pairs of real taxonomic class distributions. If it proves possible to obtain nearly the same results for real and artificial data for a large variety of situations then one would have more confidence that the method is capable of adequate simulation of real distributions for the present purposes. It is of crucial importance that there obtain a positive result here especially since it is not known how the unaccounted for degrees of freedom are absorbed by the method. In intuitive language, the suspicious possibility which must be dismissed is that the method cannot generate onerous distributions like some or many or most real ones are and the Monte Carlo study of a taxometric method with gentle and tame distributions is not sufficiently tough and is, therefore, unilluminating.

It might be possible to resolve the above problem in the following way. Suppose some real data are analyzed by a taxometric theory and a taxonomy is detected with all consistency tests passed. We then classify individuals into, say, two taxonomic classes, denoted by s and n, and determine the misclassification rates, R_1 and R_2, by a method

given In Golden and Meehl (p. 31ff, 1973a). The interitem covariances for each classification group, c_{1ij} and c_{2ij}, are given by

$$c_{1ij} = R_1 c_{sij} + (1 - R_1)\, c_{nij} + R_1(1 - R_1)(p_{si} - p_{ni})(p_{sj} - p_{nj})$$

$$c_{2ij} = R_2 c_{sij} + (1 - R_2)\, c_{nij} + R_2(1 - R_2)(p_{si} - p_{ni})(p_{sj} - p_{nj})$$

which can be solved for the taxonomic class covariances c_{sij} and c_{nij} once the p_{si}'s and p_{ni}'s have been obtained by simultaneously solving the pair of equations

$$p_{1i} = R_1\, p_{si} + (1 - R_1)\, p_{ni} \quad \text{and}$$

$$p_{2i} = R_2\, p_{si} + (1 - R_2)\, p_{ni}\ .$$

Using the resulting intra-taxonomic class correlation values, we use the random number generator method to generate a sample of the same size which is then analyzed by the taxometric method. If the consistency test and parameter estimation results are sufficiently close to the real data ones we need not doubt the random number generator method at least in this instance. Generating several such artificial samples would help determine how much sampling error must be considered.

Even though a detected taxonomic situation is simulated very well, the detection still could be inaccurate or spurious. Adequate simulation is evidently just a necessary condition for accurate and non-spurious detection.

In summary, the Monte Carlo method is resorted to for two quite different procedures. First, it is used for the approximation of the functions relating parameter estimate errors and consistency function values. Second, it is used to simulate the estimated taxonomic situation resulting from analysis of real data by a taxometric

method. Each of the two Monte Carlo procedures result in requirements, which are considered as *necessary* for the existence of a real taxonomy with parameter values as estimated by a taxometric method.

REFERENCES

Abramowitz, M. A. and Stegan, I. A. (Eds.) (1964). *Handbook of mathematical functions*. Washington, D. C.: U. S. Government Printing Office.

Anderson, T. W. (1957). *An introduction to multivariate statistical analysis*. New York: Wiley and Sons.

Golden, R. R. and Meehl, P. E. (1973a). *Detecting latent clinical taxa, IV: An empirical study of the maximum covariance method and the normal minimum chi-square method using three MMPI keys to identify the sexes*. Reports from the Research Laboratories of the Department of Psychiatry, University of Minnesota, Report Number PR-73-2, Minneapolis, MN.

Golden, R. R. and Meehl, P. E. (1973b). *Detecting latent clinical taxa, V: A Monte Carlo study of the maximum covariance method and associated consistency tests*. Reports from the Research Laboratories of the Department of Psychiatry, University of Minnesota, Report Number PR-73-3, Minneapolis, MN.

Golden, R. R., Tyan, S. H., and Meehl, P. E. (1974a). *Detecting latent clinical taxa, V I : Analytical development and empirical trials of the consistency hurdles theory*. Reports from the Research Laboratories of the Department of Psychiatry, University of Minnesota, Report Number PR-74-4, Minneapolis, MN.

Golden, R. R., Tyan, S. H. and Meehl, P. E. (1974b). *Detecting latent clinical taxa, VII: Analytical development and empirical and artificial data trials of the multi-indicator, multi-taxonomic class maximum likelihood normal theory*. Reports from the Research Laboratories of the Department of Psychiatry, University of Minnesota, MN.

Kendall, M. G. (1952). *The advanced theory of statistics*, Vol. I. (5th ed.) New York: Hafuer.

McNemar, Q. (1962). *Psychological statistics*. New York: Wiley and Sons.