Golden, R.R., & Meehl, P.E. (1973) *Detecting Latent Clinical Taxa, V: A Monte Carlo Study the Maximum Covariance Method and Associated Consistency Tests* (Report No. PR-73-3) Minneapolis, MN: Research Laboratories of the Dept. of Psychiatry, University of Minnesota

Reports from the Research Laboratories
of the Department of Psychiatry University of Minnesota

Detecting Latent Clinical Taxa, V:
A Monte Carlo Study of the Maximum Covariance Method and Associated Consistency Tests[1]

Robert R. Golden and Paul E. Meehl

Report Number PR-73-3 December 1973

TABLE OF CONTENTS

[Pagination in this digital version differs from original printing]

[1] This research was supported in part by grants from the Psychiatry Research Fund, the National institute of Mental Health, Grant Number MH 24224, and the Schizophrenia Research Program, Scottish Rite, Northern Masonic Jurisdiction, U.S.A.

I. *Introduction*

The primary purpose of this investigation was to study a variety of artificial data samples to get a rough idea of the kind of real dichotomous latent taxonomies for which the maximum covariance method (Meehl, 1965; Meehl, 1968), when used with indicators such as three MMPI keys, is capable of adequate detection and those for which it is not. The term "adequate detection'' is used to mean that the most important parameters of a dichotomous latent situation such as base-rate, valid-positive and valid-negative rates, indicator means and standard deviations are estimated accurately enough. Although the required degree of accuracy will vary from one substantive study to the next and even from one investigator to the next, there can be general agreement as to when estimates are totally off the mark and when estimate accuracy is high enough not to be of any real concern in any actual substantive application in personality measurement. Results that lie in the middle ground are regarded as suggesting further empirical and Monte Carlo study of the method.

There are two main reasons why a method gives incorrect results:

(1) the sample size is inadequate and/or

(2) the assumptions of the method are not adequate approximations to the real situation.

An important consideration is that some assumptions are more robust than others. That is, the robust assumption is one that does not perpetrate substantial error in estimates of the important parameters even though it appears to be an inadequate approximation of the actual situation. It will be shown, for example, that the model's most worrisome assumption of intra-taxon independence appears not to be a real matter of concern when the two intra-taxon covariances are approximately equal although greatly different from zero. On the other hand, substantially unequal intra-taxon covariances will be seen to be a matter of genuine concern.

II. *The Consistency Test*

A second purpose of the study was to get a rough idea of how a few consistency tests (see PR-65-2, especially pp. 24-34) [1] might be used to determine whether the results of the method (the estimates of the latent parameters) are to be taken seriously or instead regarded as probably totally erroneous, so best to be totally ignored and thrown away. Again, the middle ground

[1] [All cites to previous research reports refer to page numbers in original paper copies; digital posted versions usually have different pagination and must be searched for specific content.]

results indicate further detailed study. In short, a psychometric method is itself a mathematical model and, therefore, really another theory (see Brodbeck [1963] for the equivalence of the terms "model" and "theory") and it is the purpose of the consistency test to test the appropriateness of the psychometric theory in terms of various relationships observed in the real data which is purportedly of substantive interest. In any psychometric theory it would be possible to derive (from the given assumptions) a number of relationships involving the latent and observed parameters. Those relationships not used for estimation of the latent parameters could be used as consistency tests. The degree of consistency of the psychometric theory with the substantive data is increased roughly by (1) increasing the number of consistency tests, (2) using minimally dependent tests in the sense that they are derivable from different major assumptions (or different subsets of the same), (3) using the consistency tests of maximal power—especially for the assumptions which are known to be of questionable robustness (or worse, known to have very meager robustness) and, of course, (4) increasing the degree to which the consistency test formulae are satisfied. Ideally, the discrepancies in (4) would be smaller than, say, one probable (sampling) error although this requirement is undoubtedly far too tough; that is, too many good substantive theory-psychometric theory combinations would be refuted. This result would then be the exact opposite of the current prevailing situation where statistical hypothesis tests are largely relied upon with concern only for α-type errors and not β-type errors (Morrison & Henkel, 1970).

III. *Generation of Multivariate Artificial Data*

The maximum covariance model requires at least three keys or scales of a quasi-continuous nature. By this is meant, for example, that a sixty-item key is "more continuous" than a five-item key. For a number of reasons, it appears that MMPI keys should be over ten items long for the usual taxonomic situation and one unpublished study indicates that MMPI keys should not be longer than 20-25 items. In the present study, three simulated twenty-item keys are used.

It would be nice to use the most general analytical multivariate distributions within each taxon since multi-personality-measure distributions are of unknown complexity. What is actually known about such distributions is that it is usually true (with clear exceptions) that (1) the functional relationship between two measures is usually adequately approximated by a linear one and (2) an indicator (marginal) distribution is of a shape that is roughly normal within a taxon. If

one were only concerned with univariate (intra-taxa) distributions then he would be behooved to use the more general Pearson distribution (see Kendall, 1943-46); however, the generalization of the univariate Pearson distribution just to the bivariate case results in unwieldy complications. Such an attempt was essentially successful as a purely mathematical exercise by the mathematician Van Uven (1947), but a review of this work shows that it is not of practical value due to the large number of special cases resulting from terrible analytical complexities.

Although a few attempts at developing a rationale for generating general multivariate artificial data distributions are promising at the theoretical level it was regretfully decided to put this difficult problem aside temporarily and use intra-taxa multivariate normal distributions for this initial investigation.

Multivariate normal distributions probably come close to approximating most real distributions in personality inventory key-indicator study in view of the plausibility of bivariate linearity and marginal normality as mentioned above.

The multivariate normal distributions were generated by use of a univariate normal generator in the following wav. Let $\underset{n\times 1}{Z} = (z_1, z_2, ..., z_n,)$ be an *n*-tuple random vector such that each z_i is a standardized random variable (with zero mean and unit variance) and let the covariance between z_i and z_j be σ_{ij} for each *ij* pair, if

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

This is an approximation to the hand-drawn symbol used by Golden:

is the given covariance matrix and the z_i are distributed multivariate normally and standardized then these conditions can be written as $\underset{n\times 1}{Z} \overset{d}{=} N(0, \underset{n\times n}{\sum})$. Let $\underset{n\times 1}{Y} = (y_1, y_2, ..., y_n)$ be a set of such variates except where $\sum$ = I the identity matrix. Suppose that $\underset{n\times 1}{Z} = \underset{n\times n}{T} \cdot \underset{n\times 1}{Y}$, where *T* is an unknown transformation matrix. Thus, $\underset{n\times n}{\sum} = \mathrm{E}\left(\underset{n\times 1}{Z} \cdot \underset{1\times n}{Z'}\right) = \mathrm{E}\left(T \cdot Y\right) \cdot \left(TY\right)' = \mathrm{E}\left(TY \cdot Y'T\right) = T\mathrm{E}\left(YY'\right)T'$; since $\mathrm{E}\left(YY'\right) = \mathrm{I}$, we have, $\underset{n\times n}{\sum} = \underset{n\times n}{T} \cdot \underset{n\times n}{T'}$ and it is seen that *T* is the matrix such that the product of it and its transpose is the given $\sum$. Well known methods exist for solving the latter equation for *T*.

IV. *The Maximum Covariance Method*

The method is given in Section 1, pp. 2-7 of PR-68-4 as a revision of the original method given in Section 3, pp. 10-12 of PR-65-2. An empirical trial of the method is reported in PR-73-2, where the method is developed from precisely stated assumptions. An outline of the method is given below in terms of assumptions which are slightly different from those of the latter above report; a change is made because multivariate normal distributions partially satisfy the first assumption as formerly stated.

A. Let w, x, and y be three indicators such that w is the input indicator and x and y are the output indicators. The latent taxa distributions on the input indicator are estimated by use of manifest relationships between the two output variables.

B. The covariance between x and y for any interval of w is given by

$$\mathrm{cov}_w(x,y) = p_w\,\mathrm{cov}_{rw}(x,y) + q_w\,\mathrm{cov}_{lw}(x,y) + p_w\,q_w\,\Delta\bar{x}_w\Delta\bar{y}_w \qquad [1]$$

where p_w is the proportion of individuals in w interval that are members of the right* taxon,

q_w is the corresponding left taxon proportion ($p_w + q_w = 1$),

$\mathrm{cov}_{rw}(x,y)$ is the manifest conditional covariance between x and y for the right taxon members in interval w,

$\mathrm{cov}_{lw}(x,y)$ is the corresponding left taxon covariance,

$\Delta\bar{x}_w$ is the mean on x for the right taxon members in interval w less that for the left taxon, and

$\Delta\bar{y}_w$ is the corresponding mean difference on y.

C. Under the assumptions

A_1: $\mathrm{var}_r = \mathrm{var}_l = \mathrm{var}$, where var is the common within taxon variance on w,

A_2: $\mathrm{cov}_{rw}(x,y) = \mathrm{cov}_{lw}(x,y) = 0$ for all w, and

A3: $\mathrm{cov}_r(x,y) = \mathrm{cov}_l(x,y) = \mathrm{cov}(x,y)$ (that is, the total within taxa covariances are equal),

---

* The taxon with the highest scores on each of the input indicators (on the right side of the histogram) will be called the right taxon; the other taxon will be called the left taxon.

it follows that for multivariate normal distributions**

$$\Delta\overline{x}_w = \frac{\mathrm{cov}(x,y)}{\mathrm{var}_w}\left(w-\overline{w}_{\mathrm{r}}\right)+\overline{x}_{\mathrm{r}}-\frac{\mathrm{cov}(x,y)}{\mathrm{var}_w}\left(w-\overline{w}_{\mathrm{l}}\right)+\overline{x}_{\mathrm{l}}$$

$$=\frac{\mathrm{cov}(x,y)}{\mathrm{var}_w}\left(\overline{w}_{\mathrm{l}}-\overline{w}_{\mathrm{r}}\right)+\overline{x}_{\mathrm{r}}-\overline{x}_{\mathrm{l}}$$

which is constant for all w and, likewise, that

$$\Delta\overline{y}_w = \frac{\mathrm{cov}(x,y)}{\mathrm{var}_w}\left(\overline{w}_{\mathrm{l}}-\overline{w}_{\mathrm{r}}\right)+\overline{y}_{\mathrm{r}}-\overline{y}_{\mathrm{l}}$$

is constant for all w. Since $\Delta\overline{x}_w\Delta\overline{y}_w$ = k (a constant) for all w, it follows that $\max\{\mathrm{cov}_w(x,y)\}$ occurs in the hitmax interval (where $p_w = q_w = ½$) and the frequency distributions intersect) and is equal to the latent quantity $¼\,\Delta\overline{x}_w\Delta\overline{y}_w = ¼$k.

D. Also, it follows that

$$p_w^2 - p_w + \frac{\mathrm{cov}_w(x,y)}{\max\left\{\mathrm{cov}_w(x,y)\right\}} = 0\,, \qquad [2]$$

a quadratic with p_w and q_w as two roots. In other words, the latent frequency distributions on w for each taxon can now be estimated. From these, the latent taxa means, standard deviations, base-rates and any other marginal distribution parameters are estimated.

E. With three indicators the roles of input and output can be interchanged to produce three different arrangements as shown below.

input indicator	output indicator
key 1	key 2, key 3
key 2	key 1, key 3
key 3	key 1 , key 2

F. Strengthening the assumptions A_2 and A_3 even further to

A_4: the indicators are independent in the strongest sense within taxa

will allow for classification of individuals. That is, for any three intervals w, x, and y, on the

** This equation originally incorrectly typed as

$$\Delta\overline{x}_w = \frac{\mathrm{cov}(x,y)}{\mathrm{var}}\left(w-\overline{w}_{\mathrm{r}}\right)+\overline{x}_{\mathrm{r}}-\frac{\mathrm{cov}(x,y)}{\mathrm{var}_w}\left(w-\overline{w}_{\mathrm{l}}+\overline{x}_{\mathrm{l}}\right)$$

three keys, the density (the proportion of the individuals in the taxon with the scores w, x, and y), $\phi(w, x, \text{and } y)$, is equal to $\phi(w)\phi(x)\phi(y)$ where, for example, $\phi(w)$ is the taxon density for score w on key w. Then the probability that an individual is a member of the right taxon given by a vector of key scores (w, x, y) is

$$\Pr\left(\text{right taxon}|w,x,y\right) = \frac{P\phi_{\text{r}}}{P\phi_{\text{r}} + Q\phi_{\text{l}}} = \frac{P\phi_{\text{r}w}\phi_{\text{r}x}\phi_{\text{r}y}}{P\phi_{\text{r}w}\phi_{\text{r}x}\phi_{\text{r}y} + Q\phi_{\text{l}w}\phi_{\text{l}x}\phi_{\text{l}y}}$$

In this equation (and x,y,z in "where P" definitions following) the y (or w subscript was replaced with z, misnaming the 3 indicators as w,x,z or x,y,z instead of the w,x,y being used here. All have been changed to continue using indicators w,x,y.

where P is the base-rate for the right taxon

Q (= 1 – P) is the base-rate for the left taxon

$\phi_{\text{r}} = \phi(w, x, y)$ for the right taxon

$\phi_{\text{l}} = \phi(w, x, y)$ for the left taxon, and

$\phi_{\text{r}x}$ = the right taxon density function for indicator x, for example.

Then if the total misclassification is to be minimized it can be shown that the required classification rule is:

"Classify as 'right' if Pr('right'|w, x ,y) > .5, and classify as 'left' otherwise."

The base-rate was estimated for each of the three keys giving close, but of course, somewhat different results. For use in the classification formula the simple arithmetic average of the three estimates was used. The estimated taxon density functions were determined directly from the corresponding estimated frequency functions.

As mentioned above these assumptions are sufficient but not necessary conditions for the method. Weaker sufficient conditions than these exist but the stated assumptions are more easily analyzed in terms of multivariate normal distributions.

V. *The Robustness and Power of the Method*

Various sets of parameter values were chosen so the error of estimation of the parameters could be studied as a function of systematic variation of sample size (N = 1000, 800, 600 and 400), of base-rates (.5, .6, .7, .8, and .9), of separation between the two taxa means (2, 3/2, 1, 1/2 sigma units), of different taxa variance ratios (11/10, 4/3, 5/3, 3), of different within-taxon correlations (.1, .3, .5, and .8), and of different within-taxa correlation ratios (0, 4). The variation in taxa variance ratios tested the robustness with respect to assumption A_1; the variation of the within-taxa covariance ratios tested the robustness with respect to A_3; and the variation of the within-taxa correlations tested the classification method robustness with respect to A_4. The

following result is used in regard to analyzing robustness with respect to the zero within-interval-within-taxon covariance assumption A_2:

Theorem: When $\text{cov}(x,y) = 0$, $\text{cov}(x,w) = 0$, and $\text{cov}(y,w) = 0$ for each taxon, then it follows that $\text{cov}_{lw}(x,y) = \text{cov}_{rw}(x,y) = 0$ for all w.

Proof: This result follows under the conditions (a) constant $\bar{x}_{rw}$ and/or $\bar{y}_{rw}$ (and constant $\bar{x}_{lw}$ and/or $\bar{y}_{lw}$) for all w and (b) $\text{cov}_{rw}(x,y) \geq 0$, $\text{cov}_{lw}(x,y) \geq 0$ for all w. It would appear that these conditions would generally be met if $\text{cov}_r(x,y) = \text{cov}_l(x,y) = 0$,

If the distribution is assumed to be trivariate normal, then the variances and covariances of the conditional distributions are not a function of the value of the fixed variable. From results by Anderson (1958, p. 28) it can be shown that

$$\text{cov}\left(x, y \mid w\right) = \text{cov}\left(x, y\right) - \frac{\text{cov}\left(x, w\right)\text{cov}\left(y, w\right)}{\text{var}\left(w\right)}$$

which is zero when the three covariances on the right side are zero. When these three covariances are not zero, the formula allows for the calculation of the amount of departure from the condition of assumption A_2. In the case where $\text{cov}(x,y) = \text{cov}(x,w) = \text{cov}(y,w) = rs$ where r and s are the common between intra-taxon correlation and the common intra-taxon standard deviation, respectively, then

$$\text{cov}\left(x, y \mid w\right) = rs - \frac{rsrs}{s^2} = r\left(s - r\right)$$

for all w.

Thus we see that only when the intra-taxon correlations are systematically varied from zero, the robustness with respect to assumption A_2 is examined.

For all of the parameter set values certain things were kept constant for this study. Three indicators, integerized to a range of 0 to 20 such as those of personality keys, were always used in the three role combinations given above and the parameters were given the same values for each indicator. For each set of parameter values, each of twenty-five different independent random samples were generated and each used as data for the calculations of the method.

It should be noted that the manifest mixed group covariance curve was always smoothed around the maximum by the use of a least-squares fitting parabola, although experimentation has since shown the procedure could have been deleted without markedly affecting the method's general level of estimation accuracy and taxonomic detection power.

The various sets of parameter values are described in Table 1, and the summary of the parameter estimates is given in Table 2. The parameter estimates for a sample were regarded as accurate enough if the base-rates and hit-rate estimates were within .10 of the true values, and the means and standard deviation estimates were within one interval of the true values.

First, different total sample sizes of 1000, 800, 600 and 400, for $var_l = var_r = \sigma$, a difference between the taxa means of 2σ, $P = .5$, and zero intra-taxon correlations, each gave average errors of .01 (2%) in the estimation of P, less than $¼\sigma$ (this is ½ of an indicator interval) in the estimation of the taxa means and standard deviations.

Second, different base-rates of .6, .7, .8 and .9 for N (total sample size) = 1000, $\sigma_r = \sigma_l = \sigma$, $\mu_r - \mu_l = 2\sigma$, and zero intra-taxon correlations, gave corresponding average errors of .03, .04, .02 and .60 in the estimation of the base-rate and average errors of less than $⅜\sigma$, $½\sigma$, $½\sigma$, $½\sigma$, in the estimation of the means and standard deviations.

Third, different $\mu_r - \mu_l$ separations of $½\sigma$, 1σ, and $½\sigma$, for $N = 1000$, $\sigma_r = \sigma_l = \sigma$, $P = .5$ and zero intra-taxon correlations, gave average errors of .01 in the estimation of P and less than $¼\sigma$ in the estimation of the means and standard deviations.

Fourth, different standard deviation ratios (σ_r/σ_l) of 11/10, 4/3, 5/3 and 3 for $N = 1000$, $\mu_r - \mu_l = ½(\sigma_r + \sigma_l)$, $P = .5$ and zero intra-taxon correlations, gave averaqe errors of .02, .03, .08 and .14 in the estimation of P and average errors less than $¼\sigma$, $¼\sigma$, $¼\sigma$, $½\sigma$ in the estimation of the means and standard deviations.

Fifth, different intra-taxon correlations of .1, .3, .5 and .8 for $N = 1000$, $\mu_r - \mu_l = 2\sigma$, $\sigma_r = \sigma_l = \sigma$, and $P = .5$ gave average errors of .01 in the estimation of P and $¼\sigma$, $¼\sigma$, $½\sigma$, and 1σ in the estimation of the means and standard deviations.

More briefly, then, the method requires the following parameter boundaries in order to work well enough for most personality measurement work; i.e. base-rates accurate to within .10 and indicator means and standard deviations accurate to within $½\sigma$ (an intervaI width) :

a) sample size ≥ 400,
b) base-rates not disproportionate more than (.2, .8),
c) separation of means ≥ 1.0σ,
d) standard deviation ratio < 1.7,
e) intra-taxon correlations ≤ .5. and

f) the difference between the two corresponding intra-taxon correlations < .4.

The above conditions are, of course, necessary but not sufficient conditions for the method to work well. The only stringent condition of these would appear to be (f). Further study of this, condition will be given in a forthcoming report on an iterative modified version of the maximum covariance method, where the within taxa covariance assumptions are relaxed.

TABLE 1
Description of Sample Sets

set	variable	N	P	$\overline{x}_l$	$\overline{x}_r$	S_l	S_r	Δ	S_r/S_l	r	
1.1	N	1000	.5	8	12	2	2	2	1	0	*
1.2		800	.5	8	12	2	2	2	1	0	*
1.3		600	.5	8	12	2	2	2	1	0	*
1.4		400	.5	8	12	2	2	2	1	0	*
2.1	P	1000	.6	8	12	2	2	2	1	0	*
2.2		1000	.7	8	12	2	2	2	1	0	*
2.3		1000	.8	8	12	2	2	2	1	0	*
2.4		1000	.9	8	12	2	2	2	1	0	
3.1	Δ	1000	.5	9	12	2	2	1.5	1	0	*
3.2		1000	.5	10	12	2	2	1	1	0	*
3.3		1000	.5	11	12	2	2	.5	1	0	
3.4		1000	.5	12	12	2	2	0	1	0	
4.1	S_r/S_l	1000	.5	8	12	1 .9	2.1	2	1.1	0	*
4.2		1000	.5	8	12	1 .7	2.3	2	1.3	0	*
4.3		1000	.5	8	12	1.5	2.5	2	1.7	0	*
4.4		1000	.5	8	12	1	3	2	3	0	
5.1	r	1000	.5	8	12	2	2	2	1	.1	*
5.2		1000	.5	8	12	2	2	2	1	.3	*
5.3		1000	.5	8	12	2	2	2	1	.5	*
5.4		1000	.5	8	12	2	2	2	1	.8	
										r_l/r_r	
6.1	N	1000	.8	8	12	2	2	2	1	.5/.125	
6.2	$r_l/r_r = 4$	800	.8	8	12	2	2	2	1	.5/.125	
6.3		600	.8	8	12	2	2	2	1	.5/.125	
6.4		400	.8	8	12	2	2	2	1	.5/.125	

N: sample size
P: base-rate of the right-taxon
$\overline{x}_l$: mean of the left taxon on each indicator
$\overline{x}_r$: mean of the right-taxon on each indicator
S_l: standard deviation of the left taxon on each indicator
S_r: standard deviation of the right taxon on each indicator
Δ: $(\overline{x}_r - \overline{x}_l)/S$ where $S = (S_l + S_r)/2$
r: intra-taxon correlation between indicator pairs
*: parameter estimates judged as accurate

TABLE 2
Average True and Estimated Parameter Values

Set	Variable	H	$\hat{H}$	P_1	$\hat{P}_1$	h_1	$\hat{h}_1$	p_1	$\hat{p}_1$	$\bar{X}_1$	$\hat{\bar{X}}_1$	S_1	$\hat{S}_1$
	N												
*1.1	1000	.95±.01	.89±.01	.50	.49±.02	.96±.01	.89±.01	.94±.02	.89±.01	8.00	8.49±.12	2.00	2.44±.07
*1.2	800	.95±.01	.89±.01	.50	.49±.02	.96±.01	.90±.01	.94±.02	.89±.01	8.00	8.47±.13	2.00	2.45±.09
*1.3	600	.95±.01	.90±.01	.50	.49±.02	.96±.01	.90±.01	.94±.02	.89±.01	8.00	8.46±.14	2.00	2.46±.09
*1.4	400	.95±.01	.90±.01	.50	.50±.02	.96±.02	.90±.02	.94±.02	.89±.02	8.00	8.51±.23	2.00	2.48±.13
	P												
*2.1	.6	.95±.01	.89±.01	.60	.57±.02	.96±.01	.87±.01	.91±.02	.88±.01	8.00	8.32±.11	2.00	2.41±.10
*2.2	.7	.95±.01	.89±.01	.70	.66±.03	.95±.02	.82±.03	.88±.03	.86±.01	8.00	8.25±.14	2.00	2.40±.07
*2.3	.8	.95±.01	.89±.01	.80	.79±.02	.89±.06	.71±.04	.88±.06	.79±.04	8.00	8.26±.18	2.00	2.40±.09
2.4	.9	.27±.24	.86±.21	.90	.30±.29	.93±.17	.81±.26	.16±.16	.93±.21	8.00	5.06±2.1	2.00	1.39±.63
	Δ												
*3.1	1.5	.89±.01	.84±.01	.50	.49±.03	.91±.02	.85±.02	.88±.02	.84±.02	9.00	9.33±.16	2.00	2.28±.08
*3.2	1.0	.78±.02	.77±.03	.50	.50±.09	.79±.10	.78±.07	.78±.06	.76±.05	10.00	10.20±.32	2.00	2.14±.11
3.3	0.5	.60±.03	.79±.09	.50	.50±.15	.64±.20	.78±.15	.62±.06	.78±.12	11.00	10.80±.66	2.00	1.97±.22
3.4	0.0	.50±.02	.78±.18	.50	.47±.16	.52±.24	.74±.23	.50±.02	.82±.19	12.00	10.97±.83	2.00	1.77±.23
	s_r/s_1												
*4.1	1.1	.95±.01	.89±.01	.50	.48±.02	.95±.01	.89±.01	.94±.02	.89±.01	8.00	8.48±.10	1.90	2.38±.08
*4.2	1.3	.95±.01	.88±.01	.50	.47±.02	.95±.02	.87±.01	.95±.02	.89±.01	8.00	8.49±.10	1.70	2.23±.08
*4.3	1.7	.94±.02	.85±.02	.50	.42±.02	.95±.02	.83±.02	.94±.03	.89±.02	8.00	8.52±.19	1.50	2.13±.15
4.4	3.0	.65±.17	.69±.27	.50	.36±.19	.83±.30	.69±.30	.69±.20	.73±.29	8.00	8.91±.69	1.00	2.32±.59
	r												
*5.1	.1	.93±.01	.87±.01	.50	.49±.02	.94±.02	.80±.02	.93±.02	.80±.02	8.00	8.56±.10	2.00	2.51±.06
*5.2	.3	.90±.01	.84±.01	.50	.4 ±.02	.90±.04	.74±.03	.89±.04	.74±.02	8.00	8.75±.17	2.00	2.60±.10
*5.3	.5	.88±.01	.80±.02	.50	.50±.03	.90±.03	.81±.02	.87±.02	.79±.03	8.00	8.99±.25	2.00	2.69±.12
5.4	.8	.84±.01	.72±.04	.50	.50±.03	.88±.04	.76±.04	.82±.05	.71±.05	8.00	9.37±.37	2.00	2.81±.10
	$N(r_1/r_r=4)$												
6.1	1000	.52±.30	.69±.36	.80	.66±.22	.80±.27	.52±.35	.49±.36	.66±.38	8.00	9.28±2.4	2.00	2.22±.58
6.2	800	.55±.24	.76±.24	.80	.67±.15	.81±.28	.62±.28	.43±.29	.81±.30	8.00	9.30±3.0	2.00	2.27±.63
6.3	600	.59±.28	.73±.28	.80	.64±.22	.81±.26	.55±.30	.52±.34	.74±.33	8.00	9.65±2.2	2.00	2.21±.74
6.4	400	.56±.32	.73±.33	.80	.64±.22	.83±.19	.58±.32	.50±.35	.70±.36	8.00	9.50±3.2	2.00	2.29±.91

H: Observed mean overall hit-rate

P_1: Actual base-rate of left taxon

h_1: Actual proportion of left taxon correctly identified

p_1: Actual proportion of left taxon predictions which are correct

$\bar{X}_1$: Actual mean of left taxon

S_1: Actual standard deviation of left taxon

^: Denotes estimate

±: Number after ± is standard deviation of estimates

*: parameter estimates are judged to be accurate

VI. *Parameter Estimation Bias*

The results show that the means tend to be estimated as too close together. Also, the variances are nearly always too large. Both these results can be explained as follows. When the within-taxon between indicator covariances are zero it is true that $\mathrm{cov}_w(x,y) = p_w q_w k$, for all w, where k is estimated by $\max\{\mathrm{cov}_w(x,y)\}$, and $p_w = \dfrac{1 \pm \sqrt{1-\left(4\mathrm{cov}_w\left(x,y\right)/k\right)}}{2}$ where the minus sign is used for w less than the hitmax cut (where $\mathrm{cov}_w(x,y)$ is a maximum) and the plus sign for w greater than the hitmax cut. While the sample estimate of $\mathrm{cov}_w(x,y)$ is unbiased, it is clear that k will tend to be overestimated since it is a sample maximum. The error in p_w, Δp_w, caused by the error in k, Δk, can be estimated by

$$\Delta p_w = \frac{\pm \partial p_w}{\partial k} \qquad \text{where} \quad \frac{\partial p_w}{\partial k} = \frac{\mathrm{cov}\left(x,y\right)}{k^2\sqrt{\dfrac{1-4\mathrm{cov}\left(x,y\right)}{k}}} \; .$$

Monte Carlo study of the magnitude of Δk would allow one to correct the estimates of p_w so as to be more nearly unbiased. Presently, p_w is biased to be large for values to the right of hitmax or for most of those values of p_w which are substantially greater than zero. Since

$$\frac{\partial^2 p_w}{\partial \mathrm{cov}\left(x,y\right)\partial k} = \frac{1}{k^2\left[1-\dfrac{4\mathrm{cov}\left(x,y\right)}{k}\right]^{1/2}} + \frac{\mathrm{cov}\left(x,y\right)}{\left[1-\dfrac{4\mathrm{cov}\left(x,y\right)}{k}\right]^{3/2}}$$

is positive, we see that $\dfrac{\partial p_w}{\partial k}$ is a monotonically increasing function of $\mathrm{cov}(x,y)$, and since p_w is directly proportional to $\dfrac{\partial p_w}{\partial k}$ it is clear that $\Delta \mathrm{p}_w$ will be a monotonically decreasing function of w (for w greater than hitmax). This will tend to cause the right taxon distribution to be biased toward the left and similarly the left one to the right. Thus, the means are each biased toward the middle and the variances biased to be too large. Also, the base-rate for the right taxon will be biased too large since the Δp_w are positive and the base-rate for the left will be too small since the $\Delta\sigma_w$'s were calculated from $\sigma_w = (1 - p_w)$; this is observed to generally be the case.

VII. *The "Total Covariance" Consistency Test* [T_1]

The covariance mixture equation can be written for the total mixed group as

$$\mathrm{cov_m}(x,y) = P\,\mathrm{cov_r}(x,y) + Q\,\mathrm{cov_l}(x,y) + PQK \qquad [3]$$

where P is the base-rate of the right taxon

Q is the base-rate of the left taxon, and

K is the product of the differences in taxa means on x and y.

It follows that if $\mathrm{cov_r}(x,y) = \mathrm{cov_l}(x,y) = 0$ as A_3 requires, then the quantity

$$T_1 = \widehat{\mathrm{cov}}_{\mathrm{m}} - \hat{P}\hat{Q}\hat{K} \qquad [4]$$

where the carat denotes parameter estimates, can be expected to be close to zero if the parameter estimates are accurate. By considering the differential of T_1 we have

$$dT_1 = \frac{\partial T_1}{\partial \mathrm{cov}_{\mathrm{m}}(x,y)}\, d\,\mathrm{cov}_{\mathrm{m}}(x,y) + \frac{\partial T_1}{\partial P}\, dP + \frac{\partial T_1}{K}\, dK$$

and it is shown that

$$\Delta T_1 \le \Delta\,\mathrm{cov}_{\mathrm{m}}(x,y) + PQ\Delta\bar{x}\Delta(\Delta\bar{y}) + PQ\Delta\bar{y}\Delta(\Delta\bar{x}) + (1-2P)\Delta\bar{x}\Delta\bar{y}\Delta P \qquad [5]$$

If parameters are estimated accurately enough for practical work then

$$\frac{\Delta(\Delta\bar{y})}{s_y} \le 1/2 \quad \text{and} \quad \frac{\Delta(\Delta\bar{x})}{s_x} \le 1/2\ ,$$

when s_x and s_y are the within taxa standard deviations and $\Delta P < .1$. It can tentatively be assumed that

$$\frac{\Delta\bar{x}}{s_x} \le 2 \quad \text{and} \quad \frac{\Delta\bar{y}}{s_y} \le 2$$

if that is what the parameter estimates indicate and since larger mean separations should allow for rather simple taxonomic parameter estimation. For the method to work well enough we have shown that $P > .2$ or $(1 - 2P) < .6$. By use of Fisher's z-transformation, it can be shown that $\mathrm{cov_m}(x,y) < .64$ with a probability of more than .95. Hence, we can conclude from [5] that

$$\Delta T_1 < .64 + 1/4 \cdot 2 \cdot \frac{s_y s_x}{2} + 1/4 \cdot 2 \cdot 2 \cdot \frac{s_x s_y}{2} + .6 \cdot 2 \cdot s_x \cdot 2 \cdot s_y \cdot .1$$

or, if $s_x = s_y = s$,

$$\Delta T_1 < .64 + .74 s^2$$

As was shown above $\hat{s} = s$, so $T_1 < .64 + .74\hat{s}^2$. For the present trivariate arrangement the test can be applied three times for each sample. For the consistency test to be passed it will be required that all three values of T_1 be less than the above limit. The results of the test are given in Table 3. The test is apparently a sensitive detector of within-taxa correlations of .5 or above; this result is certainly reasonable since the test rests squarely on the assumption that these correlations are zero.

Various sets of twenty-five samples each were generated from three and four taxa with small mean separations and equal base-rates. Each of these samples failed this consistency test.

VIII. *The "Maximum Mean Difference" Consistency Test* [T_2][1]

If we consider a cut score on the input variable, then the mean of the individuals with scores above the cut can be calculated, call it $\overline{x}_{aw}$; similarly for the mean below, call it $\overline{x}_{bw}$. Then these quantities can be calculated for all values of *w* and the maximum of the difference, $\max\left\{\overline{x}_{aw} - \overline{x}_{bw}\right\}$, determined. It was argued in PR-68-4 that the maximum should occur near the hitmax cut. It can be shown that if $\overline{x}_{lw}$ and $\overline{x}_{rw}$ are each constant for all *w* then the function $\overline{x}_{aw} - \overline{x}_{bw}$ is concave downward with minima at the endpoints of the scale. It is easily seen that

$$\lim_{w \to w_{\max}} \left(\overline{x}_{aw} - \overline{x}_{bw}\right) = \overline{x}_r - \left(P\overline{x}_r + Q\overline{x}_l\right) = Q\left(\overline{x}_r - \overline{x}_l\right)$$

and

$$\lim_{w \to w_{\min}} \left(\overline{x}_{aw} - \overline{x}_{bw}\right) = P\overline{x}_r + Q\overline{x}_l - \overline{x}_l = P\left(\overline{x}_r - \overline{x}_l\right) .$$

If $T_2 = \max(\overline{x}_{aw} - \overline{x}_{bw})$ is larger than either of these minimum extrema then it must be true that

$$T_2 = \max\left(\overline{x}_{aw} - \overline{x}_{bw}\right) > 1/2\left(\overline{x}_r - \overline{x}_l\right) .$$

If the means are far enough apart so that the method provides accurate parameter estimates, then $\overline{x}_{rw} - \overline{x}_{lw} > s$, where *s* is the common standard deviation; thus $T_2 > s/2$. As would be hoped, this test correctly identified all samples of the two parameter sets where the separation in the means was ½ *s* and zero. It also incorrectly rejected 2 of the 25 samples when the separation was 1*s* and the parameters were accurately estimated.

---

[1]This consistency test subsequently became the MAMBAC (Mean Above Minus Below A Cut) procedure (Meehl & Yonce, 1994).

The parameter T_2 also has an upper limit. Since the largest value of $\max\{\bar{x}_{aw} - \bar{x}_{bw}\}$ is obtained when the taxa are totally separated, this value is clearly $\bar{x}_r - \bar{x}_l$. Thus $T_2 < \bar{x}_r - \bar{x}_l$ or $T_2 < \hat{\bar{x}}_r - \hat{\bar{x}}_l$. This test turned out to be very sensitive in the detection of more than two taxa (as did the first consistency test). When the taxon correlations were .8 the parameter estimates were quite inaccurate. This test detected each of these samples; also 7 of the samples where the correlation was .5 were incorrectly rejected,

IX. *The "Hit-Rate Proportion" Consistency Test* [T_3]

If we consider a cut on the input variable and the resulting proportion of the individuals above the cut that are correctly identified, h_{aw}, and the corresponding proportion below, h_{bw}, and let the proportion of the total number of individuals above the cut be P_{aw} and that below be P_{bw}, then the quantity $\frac{P_{aw}}{P_{bw}} - \frac{u_{aw}}{u_{bw}}$, where $u_{aw} = h_{aw} - 1/2$ and $u_{bw} = h_{bw} - 1/2$, is argued to have a minimum value near the hitmax cut in PR-68-4. It appears that for parameter estimates to be accurate it is necessary that $T_3 = \min\left\{\frac{P_{aw}}{P_{bw}} - \frac{u_{aw}}{u_{bw}}\right\} < 2$. This test correctly identified those samples where the taxa variance ratios were 3 and gave incorrect parameter estimates. A small percentage of other parameter sets producing incorrect estimates were also correctly identified. The test incorrectly identified 10 of the 23 samples where the separation in the means was 1σ and accurate estimates produced.

X. *The "Sum of the Hit-Rates" Consistency Test* [T_4]

In PR-68-4 it is arqued that $T_4 = \max\{h_{aw} + h_{bw}\}$ occurs near the hitmax cut. For the present purposes it is clear that for a cut near hitmax for any distribution with base-rates not more disproportionate than (.2, .8), $h_{aw} > ½$ and $h_{bw} > ½$, or $T_4 > 1$. If the separation between the means is only 1σ it can be shown that for normal distributions with base-rates not more disproportionate than (.2, .8) that $T_4 > 1.3$.

This test proved sensitive in the detection of base-rates more disproportionate than (.2, .8), variance ratios of 3 and correlation ratios of 4, all of which produced inaccurate parameter estimates. A very small percent of the samples where parameter estimates were accurate were incorrectly rejected.

TABLE 3

Joint frequency distribution of accurate-not accurate parameter estimates and pass-fail of each consistency test

set		# samples	consistency test							
			test 1		test 2		test 3		test 4	
			pass	fail	pass	fail	pass	fail	pass	fail
1.1	accurate	25	25	0	25	0	25	0	25	0
	not accurate	0	0	0	0	0	0	0	0	0
1.2	accurate	25	25	0	25	0	25	0	25	0
	not accurate	0	0	0	0	0	0	0	0	0
1.3	accurate	25	25	0	25	0	25	0	25	0
	not accurate	0	0	0	0	0	0	0	0	0
1.4	accurate	24	24	0	24	0	25	0	24	0
	not accurate	1	1	0	0	1	0	0	0	1
2.1	accurate	22	22	0	22	0	25	0	23	0
	not accurate	3	3	0	3	0	0	0	3	0
2.2	accurate	23	23	0	23	0	25	0	23	0
	not accurate	2	2	0	2	0	0	0	2	0
2.3	accurate	25	25	0	25	0	25	0	17	8
	not accurate	0	0	0	0	0	0	0	0	0
2.4	accurate	0	0	0	0	0	0	0	0	0
	not accurate	25	25	0	25	0	25	0	25	0
3.1	accurate	25	25	0	25	0	25	0	25	0
	not accurate	0	0	0	0	0	0	0	0	0
3.2	accurate	23	23	0	23	0	13	10	15	8
	not accurate	2	2	0	2	0	0	2	1	1
3.3	accurate	0	0	0	0	0	0	0	0	0
	not accurate	25	25	0	0	25	17	8	10	15
3.4	accurate	0	0	0	0	0	0	0	0	0
	not accurate	25	25	0	0	25	18	7	5	20
4.1	accurate	25	25	0	25	0	25	0	25	0
	not accurate	0	0	0	0	0	0	0	0	0
4.2	accurate	25	25	0	25	0	25	0	25	0
	not accurate	0	0	0	0	0	0	0	0	0
4.3	accurate	25	25	0	25	0	25	0	25	0
	not accurate	0	0	0	0	0	0	0	0	0
4.4	accurate	0	0	0	0	0	0	0	0	0
	not accurate	25	25	0	25	0	25	0	25	0
5.1	accurate	25	25	0	25	0	25	0	25	0
	not accurate	0	0	0	0	0	0	0	0	0
5.2	accurate	25	25	0	25	0	25	0	25	0
	not accurate	0	0	0	0	0	0	0	0	0
5.3	accurate	25	18	7	25	0	25	0	17	8
	not accurate	0	0	0	0	0	0	0	0	0
5.4	accurate	0	0	0	25	0	25	0	0	0
	not accurate	25	0	25	0	0	0	0	0	25
6.1	accurate	0	0	0	0	0	0	0	0	0
	not accurate	25	25	0	25	0	25	0	0	25
6.2	accurate	0	0	0	0	0	0	0	0	0
	not accurate	25	25	0	25	0	17	8	0	25
6.3	accurate	0	0	0	0	0	0	0	0	0
	not accurate	25	25	0	25	0	17	8	0	25
6.4	accurate	0	0	0	0	0	0	0	0	0
	not accurate	25	25	0	25	0	19	6	0	25

XI. *Conclusions*

As can be seen from Table 3, every sample that produced incorrect sample estimates failed at least one of the four consistency tests. Only in set 3.2 where the mean separation was 1σ were the results incorrectly rejected; it should be noted that these samples produced only marginally acceptable parameter estimates. Thus, the consistency tests worked nearly perfectly for this study. Further, it is important to discover that the single sample instance in which the consistency tests "failed" was one where good results were incorrectly rejected, rather than one of accepting erroneous parameter estimates. It appears so far that if the consistency tests say "okay", one can rely heavily on this clearance.

In PR-68-4, several more consistency tests are suggested and should be tried; those chosen for this study are only the first steps. Most likely, many better ones will be found.

This method and consistency tests should be studied with more general, or at least, different distributions than multivariate normal ones. Most of the results obtained here can be confidently used in practice only when approximate multivariate normality is obtained.

The relation between size of the consistency test parameter and size of the parameter estimate error is clearly not a dichotomous one as was used for summarizing the results in Table 3, Analytical or more detailed Monte Carlo investigation of the total relation would provide further insight via the consistency test parameters.

There is another consistency test of a different nature than those described. Suppose that a real data sample provided parameter estimates indicating that the within taxa distributions were approximately multivariate normal. This could be determined, for example, by use of goodness of fit χ^2 tests on the mixed marginal (indicator) distribution and so forth. Using the produced parameter estimates, artificial data samples of the same size could be generated and analyzed by the method. The resulting parameter estimates should show (a) stability and (b) sufficient accuracy. If either of these conditions do not obtain then the real sample results should be rejected. If both conditions do obtain but the artificial estimates differ substantially from the real ones, then the possibility of non-multivariate normality must be considered. If this consistency test method were used with the present data, then all parameter sets would be perfectly identified.

XII. *Further Comments on the Consistency Test*

The estimation procedures are complemented by "consistency tests" which describe "how well the model does fit" the data in a sense that is analogous to how the three validity keys are used with profile interpretation of the MMPI. That is, some people malinger, randomly respond, are too defensive, lie and so on to such an extent that the MMPI item responses do not serve as "indicators" of "the underlying latent phenomena" or "personality," if you will, in a manner that is consistent with the nomothetic MMPI model of pathological personality phenomena. If consistency tests prove to be worthy of consideration in psychometric model building, then it is interesting to note that this simple idea was used by clinicians long before psychometricians. It would appear that consistency testing in utilization of a mathematical model is just as obviously required and is a matter of simple common sense, just as it was to builders of the MMPI. Anyone would know that some people will randomly respond and lie and so on when taking the MMPI. But it seems that few psychologists act as if Nature could be more devious than mathematicians require.

In general, any mathematical model 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 some involve only manifest parameters. Most of the equations which are of the most immediate concern in the development of a model involve both kinds of parameters.

There are two special types of equations:

(1) the assumptions and

(2) 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 (2) from (1). 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 model 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. The resulting set of equations can be used for determinating how well the model fits the data of the real phenomena; hence can be called the

(3) "consistency equations."

If the assumptions (1) are roughly correct and the estimates of manifest parameters (obtained

directly from the data, of course) and of the latent parameters (by the calculations given by (2)) are roughly correct, then substitution of the parameter estimates into (3) will show that they roughly satisfy each equation of (3).

There are at least three sources of parameter estimation error. *First*, the assumptions (1) are always mathematical idealizations and never strictly true for real phenomena, and therefore it is clear that the estimates resulting from (2) will contain some error and therefore substitution into (3) will reveal that these estimates are not perfect solutions of (3). *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 (2)). *Third*, it can be true that the calculation method used in (2) is one that according to the underlying mathematical theory gives at best an approximate solution to the equations resulting from (1). Therefore, it is clear that at best we can only hope for approximate solutions of (3).

Theoretically, sampling error, measurement error, and "solution" error can be assessed rather directly and can be reduced to nearly any arbitrarily small size. However, "assumption" error does not seem to be of this same sort in that its size cannot be directly assessed (since an "assumption" as opposed to a "hypothesis" is by definition not directly testable) and is not reducible to nearly any arbitrarily small size by any systematic procedure as would be used in reducing (a) sampling error (increase the sample size), (b) measurement error (reliability theory, factor analysis, and item selection methods provide general guidance), or (c) solution error (for example, continue an iterative calculation procedure until convergence conditions are adequately satisfied). The basic difference between assumption error and the others is evidenced by the existence of a variety of theories to assess the latter while there is, of course, 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 model so that the parameter estimates of (2) become closer to perfect solutions of (3) would increase the verisimilitude of the model (assuming that the consistency equations are chosen correctly so as to provide sufficient testing of the fit of the model to the data). The consistency testing development might 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 (2); even the addition of "weak"

assumptions to (2) should not allow the derivability of (3),

(c) they follow as "directly" from (1) as does (2) and, in fact, might be partially interchangeable with (2),

(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 model might be revised to obtain a better fit (by pointing out the set of disparate assumptions with the aid of (a)), and

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

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

In summary, the idea of consistency testing follows that of the physicists and other natural science mathematical model builders. In any mathematical model fitting there are latent variables (not directly observable) call them $x_1, x_2, x_3, \ldots$ and manifest variables $y_1, y_2, y_3, \ldots$. According to some psychometric theory the x_i's and the y_i's are related by a set of equations $f_j\left(\tilde{x}_i, \tilde{y}_i\right) = 0$ which results from some set of assumptions $A_1, A_2, \ldots, A_n$. Let the set of equations

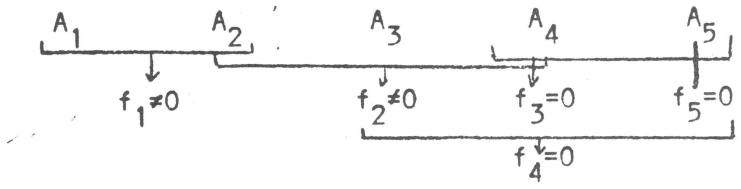
$$F = \left\{ f_j\left(\tilde{x}_i, \tilde{y}_i\right) = 0 \right\}$$

be such that the A_i are included as a subset of F. Then in all usual mathematical model developments some of those member equations of F are such that they involve both the $\tilde{x}_i$ and the $\tilde{y}_i$ and are used such that each of the $\tilde{x}_i$ are solved for in terms of the $\tilde{y}_i$ which, of course, are estimated directly from the data. This is done by whatever mathematical means one can best come up with. Consideration is usually given to the computational time (so as to decrease expense) and to the avoidance of excessive propagation of errors in the $\tilde{y}_i$ to errors in the $\tilde{x}_i$.

With enough mathematical competency it is possible to derive many <u>further</u> members of F (possibly it could become of infinite size). Each of these further member equations can each be checked to see if the already obtained approximations of the latent and manifest parameters when plugged in are approximately true. This is the consistency test. We do not know how "approximately true" they must be so we must be guided by intuition and experience and possibly the use of a higher level theory of error propagation. <u>Psychometric theories (equals "methods" equals "models") hardly ever make any attempt to do such internal checking</u>. An

exception to some extent is the Lazarsfeld and Henry (1968) discussion of the goodness of fit tests of the latent class model where it is suggested that the manifest joint compound proportions for the various higher order indicator patterns be checked against those predicted from the previously acquired estimates of the lower order proportions. Typically, the investigator of a substantive problem has a real data matrix, applies the calculations and gets his answers, say, estimates of factor loading parameters. He looks at the relative sizes and patterns; it usually makes some sense and he goes directly from there back to within his own substantive theory ballpark. But how much corroborative support has been found for the general application of the psychometric theory of factor analysis to the testing of substantive theories in a wide variety of psychological phenomena? One might resort to the total payoff so far in factor analytic studies (indicated, say, by the popularity of the method) in an area, for example, such as personality measurement as evidence that somehow all assumptions must be in pretty good agreement with the truth, generally. But Lykken (1971), for example, has taken nearly a diametrically opposite view of this particular "total payoff" so far in the personality measurement area. Factor analysis is a good example of how psychometric theory that is unsupported by development of consistency tests is still faithfully used to test substantive theories. Another good example might be internal consistency reliability theory. It is always simple to obtain an estimate of a parameter according to calculations of some method, but that doesn't mean the estimates are meaningful and accurate.

Adequately developed consistency tests such as in the example diagrammed below could help pinpoint assumptions that cannot be lived with.



Since $f_1 \neq 0$ and $f_2 \neq 0$ but $f_3 = f_4 = f_5 = 0$ (in an approximate sense) then we know that A_2 is likely to be wrong. Presumably, we would be able to alter this assumption and we would probably desire to "weaken" it. After this model alteration we could try again. In similar fashion, the data could be transformed and generally controlled by whatever means possible so that it fits the models for which we are capable of handling the mathematical problems. The physicist does both; that is, he continually alters his model to gradually fit better, and he continually alters his

data collection methods so the data are more easily fitted. Further, he does both simultaneously in an integrated fashion. The measurement psychologist does neither "model bootstrapsing" nor "data bootstrapsing." An integrated bootstrapsing is blatantly lacking since models are invented by the methodologically inclined while the data are collected by the substantively inclined, and neither type usually pays the necessary attention to what the other is really doing.

REFERENCES

Anderson, T.W. (1958) *An introduction to multivariate statistical analysis*. New York: Wiley.

Brodbeck, M. (1963) Logic and scientific method in research on teaching. In N.L. Gage (Ed.), *Handbook of research on teaching*. Chicago: Rand McNally.

Golden, R.R. and Meehl, P.E. (1973) *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* (Report No. PR-73-2). Minneapolis, MN: Research Laboratories of the Department of Psychiatry, University of Minnesota.

Kendall, M.G. (1943-46) *The advanced theory of statistics*. London: C. Griffin.

Lazarsfeld, P.R., & Henry, N.W. (1968) *Latent structure analysis*. New York: Houghton Mifflin.

Lykken, D.T. (1971) Multiple factor analysis and personality research. *Journal of Experimental Research in Personality*, *5*, 161-170.

Meehl, P.E. (1965) *Detecting latent clinical taxa by fallible quantitative indicators lacking an accepted criterion* (Report No. PR-65-2). Minneapolis, MN: Research Laboratories of the Department of Psychiatry, University of Minnesota.

Meehl, P.E. (1968) *Detecting latent clinical taxa, II: A simplified procedure, some additional hitmax cut locators, a single-indicator method, and miscellaneous theorems* (Report No. PR-68-4). Minneapolis, MN: Research Laboratories of the Department of Psychiatry, University of Minnesota.

Meehl, P.E., & Yonce, L.J. (1994) Taxometric analysis: I. Detecting taxonicity with two quantitative indicators using means above and below a sliding cut (MAMBAC procedure). *Psychological Reports, 74*, 1059-1274.

Morrison, D.E. & Henkel, R.E. (Eds.) (1970) *The significance test controversy*. Chicago: Aldine.

Van Uven, M.J. (1947) Extension of Pearson's probability distributions to two variables. *Proceedings Academie van Wetnschappen Amsterdam*, *50*, 1063-1070.