Multivariate Statistical Inference
under Marginal Structure, II

by

Leon Jay Gleser*
Purdue University

Ingram Olkin
Stanford University

Department of Statistics
Division of Mathematical Sciences
Mimeograph Series #325

July 1973

*Dr. Gleser's research sponsored in part by the Air Force Office of Scientific Research, Air Force Systems Command, USAF, under AFOSR Contract No. F44620-70-C-0066 at Johns Hopkins University and AFOSR Grant No. 72-2432 at Purdue University. The United States Government is authorized to reproduce and distribute reprints for Government purposes notwithstanding any copyright notation hereon.
Multivariate Statistical Inference under Marginal Structure, II.

by

Leon Jay Gleser

Purdue University

and

Ingram Olkin

Stanford University

1. Introduction:

A previous paper [Gleser and Olkin (1972)] dealt with statistical inference problems in the context of an experimental design in which \( k \) randomly chosen groups of individuals from the same population of individuals are asked to take different psychological tests under identical testing conditions. The \( k \) tests \((T_0, T_1), (T_0, T_2), \ldots, (T_0, T_k)\) have one subtest \( T_0 \) in common. It is desired to test whether \((T_0, T_1), (T_0, T_2), \ldots, (T_0, T_k)\) are parallel forms of the same test.

The experimental design described above is a natural one for ongoing testing programs such as the Scholastic Aptitude Test (SAT), where forms must be changed from year to year (so that new items must be introduced and validated), yet experimentation with new forms and administration of old forms are done simultaneously. The design described above might be used in a given year to validate items for future years.

Another experimental design which has potential use in ongoing testing programs is one involving a certain hierarchical structure. Because this
design is more complex than the design mentioned above, we confine our discussion to the case where 3 groups of individuals are randomly chosen from a given population of individuals and are tested under identical conditions. The three groups are given tests of the form

\[(1.1) \quad (T_0, U_1, V_1), (T_0, U_1, V_2), (T_0, U_3, V_3), \]

respectively. Here \(T_0\) is a subtest common to all three tests, \(U_1\) is common to the first two groups, and \(V_1, V_2, V_3\) are new subtests. The tests are characterized by \(r = r_0 + r_1 + r_2\) scores, \(r_0\) scores on subtest \(T_0\), \(r_1\) scores on subtest \(U_1\) or \(U_3\), and \(r_2\) scores on subtest \(V_1, V_2, V_3\).

Let \(x_0^{(g)}, x_1^{(g)}, x_2^{(g)}\) be the scores of a typical individual in the \(g\)-th group on subtests \(T_0, U_g, V_g\) respectively (where \(U_g = U_1, g = 1, 2\)). By our assumptions about the subtests, \(x_0^{(g)}\) is an \(r_0\)-dimensional (row) vector, \(x_1^{(g)}\) is an \(r_1\)-dimensional (row) vector, and \(x_2^{(g)}\) is an \(r_2\)-dimensional (row) vector. Thus,

\[x^{(g)} = (x_0^{(g)}, x_1^{(g)}, x_2^{(g)})\]

is an \(r\)-dimensional (row) vector, \(r = r_0 + r_1 + r_2\). We assume that \(x^{(g)}\) has a multivariate normal distribution with mean vector

\[(1.3) \quad \mu^{(g)} = (\mu_0^{(g)}, \mu_1^{(g)}, \mu_2^{(g)}),\]

and covariance matrix

-2-
(1.4) \[ \Sigma^g = \begin{pmatrix} \Sigma_{00}^g & \Sigma_{01}^g & \Sigma_{02}^g \\ \Sigma_{10}^g & \Sigma_{11}^g & \Sigma_{12}^g \\ \Sigma_{20}^g & \Sigma_{21}^g & \Sigma_{22}^g \end{pmatrix} \]

where the blocking of \( \mu^g \) and \( \Sigma^g \) conforms to the blocking of \( x^g \).

That is, \( \mu_{j}^g \) is \( 1 \times r_j, j = 0,1,2 \), and \( \Sigma_{jk}^g \) is \( r_j \times r_k, j,k = 0,1,2 \).

With this background in mind, we are interested in using observations obtained from the individuals tested to determine whether the 3 forms

(\( T_0, U_1, V_1 \)), (\( T_0, U_1, V_2 \)), and (\( T_0, U_3, V_3 \)) are parallel forms of the same psychological test. If the 3 forms are parallel, then the parameters \( \mu^1, \mu^2, \mu^3, \Sigma^1, \Sigma^2, \Sigma^3 \) satisfy the null hypothesis:

(1.5) \[ H: \mu^1 = \mu^2 = \mu^3 \text{ and } \Sigma^1 = \Sigma^2 = \Sigma^3 \]

If the 3 forms are not parallel, then the construction of our experimental design assures us that at least the following relationships hold among the parameters:

(1.5) \[ A: \begin{cases} \mu_0^1 = \mu_0^2 = \mu_0^3, & \mu^1 = \mu^2, \\ \Sigma^1 = \Sigma^2 = \Sigma^3, & \Sigma^1 = \Sigma^2 = \Sigma^3 \end{cases} \]

Thus, statistical verification of the hypothesis that the three forms

(\( T_0, U_1, V_1 \)), (\( T_0, U_1, V_2 \)), (\( T_0, U_3, V_3 \)) are parallel takes the form of a test
of the null hypothesis \( H \) against the alternative hypothesis \( A \).

In Section 2, the likelihood ratio test statistic for our hypothesis testing problem is derived. To carry out the test we use an asymptotic chi-square test. In Section 4, we show how our approach to deriving a test of the null hypothesis \( H \) can be used to construct statistical tests of hypothesis for the parallelism of test forms under various experimental designs related to those considered here and in the previous paper [Gleser and Olkin (1972)]. In Section 3 we provide an example in which some of the results of Section 2 are applied.

2. The Likelihood Ratio Test Statistic.

Assume that \( N \) individuals take the psychological test \( (T_0, U_g, V_g) \), \( g = 1, 2, 3 \). Let \( x_i^{(g)} = (x_{i0}^{(g)}, x_{i1}^{(g)}, x_{i2}^{(g)}) \) be the vector of scores of the \( i \)-th individual who takes the test \( (T_0, U_g, V_g) \), \( i = 1, 2, \ldots, N_g; g = 1, 2, 3 \). Under our experimental design, we may assume that the score vectors \( x_i^{(g)} \) are mutually statistically independent, \( i = 1, 2, \ldots, N_g; g = 1, 2, 3 \). In this case, we need not consider all of the data, but may reduce our consideration to the sufficient statistic

\[
(x, y) = (x^{(1)}, x^{(2)}, x^{(3)}, y^{(1)}, y^{(2)}, y^{(3)})
\]
where

\[
\bar{x}(g) = \frac{1}{N_g} \sum_{i=1}^{N_g} (x_{i0}^{(g)}, x_{i1}^{(g)}, x_{i2}^{(g)}) = (\bar{x}_0^{(g)}, \bar{x}_1^{(g)}, \bar{x}_2^{(g)})
\]

and

\[
v(g) = \begin{pmatrix}
v_{00}^{(g)} & v_{01}^{(g)} & v_{02}^{(g)} \\
v_{10}^{(g)} & v_{11}^{(g)} & v_{12}^{(g)} \\
v_{20}^{(g)} & v_{21}^{(g)} & v_{22}^{(g)}
\end{pmatrix},
\]

with

\[
v^{(g)} = \sum_{i=1}^{N_g} (x_{ia}^{(g)} - \bar{x}_a^{(g)})' (x_{ib}^{(g)} - \bar{x}_b^{(g)})
\]

\(a,b = 0,1,2; \ g = 1,2,3\). The statistics \(\bar{x}^{(1)}, \bar{x}^{(2)}, \bar{x}^{(3)}, v^{(1)}, v^{(2)}, v^{(3)}\) are mutually statistically independent. Further, \(\bar{x}^{(g)}\) has an \(r\)-variate normal distribution with mean vector \(\mu^{(g)}\) and covariance matrix \((N_g)^{-1} \Sigma^{(g)}\), and \(v^{(g)}\) has a Wishart distribution with degrees of freedom \(n_g = N - 1\) and parameter \(E(n_g^{-1} v^{(g)}) = \Sigma^{(g)}\), \(g = 1,2,3\). From the above facts the joint density function \(p(\bar{x}, v | \Sigma)\) of the sufficient statistics \((\bar{x}, v)\) can be directly obtained, and can be exhibited as a function of the unknown parameters

\[
\bar{\mu} = (\mu^{(1)}, \mu^{(2)}, \mu^{(3)}), \quad \bar{\Sigma} = (\Sigma^{(1)}, \Sigma^{(2)}, \Sigma^{(3)})
\]
To obtain the likelihood ratio test statistic (LRTS) of \( H \) versus \( A \), we could proceed ab initio to obtain the maximum of \( p(\bar{x}, V | \mu, \Sigma) \) over the parameters \( \mu, \Sigma \), both under the restrictions (1.4) on \( \mu, \Sigma \) imposed by \( H \), and under the restrictions (1.5) on \( \mu, \Sigma \) imposed by \( A \). The ratio of these maxima

\[
\lambda = \frac{\max \ p(\bar{x}, V | \mu, \Sigma)}{\max \ p(\bar{x}, V | \mu, \Sigma)}
\]

is then the LRTS for testing \( H \) versus \( A \). The null hypothesis \( H \) is rejected when \( \lambda \) is smaller than a predetermined constant \( \lambda^* \), where \( \lambda^* \) is chosen so as to give a desired level of significance \( \alpha \) for the test.

However, the determination of the LRTS is more easily accomplished by using the fact that \( \lambda \) is the product of the LRTS \( \lambda_1 \) for testing \( H \) versus the alternative

\[
(\mu_0^{(1)}, \mu_1^{(1)}) = (\mu_0^{(2)}, \mu_1^{(2)}) = (\mu_0^{(3)}, \mu_1^{(3)})
\]

(2.4) \( H* : \)

\[
\begin{bmatrix}
\Sigma^{(1)}_{00} & \Sigma^{(1)}_{01} \\
\Sigma^{(1)}_{10} & \Sigma^{(1)}_{11}
\end{bmatrix} = \begin{bmatrix}
\Sigma^{(2)}_{00} & \Sigma^{(2)}_{01} \\
\Sigma^{(2)}_{10} & \Sigma^{(2)}_{11}
\end{bmatrix} = \begin{bmatrix}
\Sigma^{(3)}_{00} & \Sigma^{(3)}_{01} \\
\Sigma^{(3)}_{10} & \Sigma^{(3)}_{11}
\end{bmatrix}
\]

and the LRTS \( \lambda_2 \) for testing \( H* \) versus \( A \) (Anderson (1958), Lemma 10.3.1). The hypothesis testing problems that give rise to \( \lambda_1 \) and \( \lambda_2 \) are of a form for which the LRTS has already been obtained [Gleser and Olkin (1972)]; thus, putting together the solutions of these two component testing problems:
immediately yields the desired LRTS $\lambda$.

2.1. The likelihood ratio test statistic $\lambda_1$.

Turning first to the test of $H$ versus the alternative $H^*$, we see by comparing (1.4) and (2.4) that under both hypotheses the subvectors $(x_{0}^{(g)}, x_{1}^{(g)})$ of scores on $(T_{0}, U_{g})$ have identical marginal distributions, $g = 1, 2, 3$. Thus, without loss of statistical generality, we can regard $(T_{0}, U_{1})$ and $(T_{0}, U_{3})$ as being identical subtests $Z = (T_{0}, U)$. Hence, we have three psychological tests $(Z, V_{1})$, $(Z, V_{2})$, $(Z, V_{3})$, and want to test whether these three tests are parallel forms of the same psychological test. In the notation of Gleser and Olkin (1972), henceforth abbreviated G-O, this last hypothesis testing problem is to test the null hypothesis $H_{m,c}$ that all three tests $(Z, V_{1})$, $(Z, V_{2})$, $(Z, V_{3})$ have identically distributed score vectors against the alternative hypothesis $H_{m',c}$, that only the score subvectors on the subtest $Z$ are identically distributed. From the results obtained in G-O, the LRTS for this problem is

$$
\lambda_1 = \left( \frac{3}{N} \prod_{g=1}^{3} \left| \frac{1}{N} v_{22.0, 1}^{(g)} \right|^{N/2} \right) \left( \frac{\left| \frac{1}{N} (I_{01} + I_{01}^{T}) (W + A) (I_{01} + I_{01}^{T})' \right|^2}{\frac{1}{N} (W + A)} \right)^{N/2},
$$

where $N = N_{1} + N_{2} + N_{3}$, $I_{p}$ is the $p \times p$ identity matrix,

$$
W = V^{(1)} + V^{(2)} + V^{(3)}, \quad x = \frac{1}{N} \sum_{g=1}^{3} \sum_{g=1}^{N} x^{(g)},
$$

$$
A = \frac{3}{N} \sum_{g=1}^{3} \sum_{g=1}^{N} (x^{(g)} - \bar{x})' (x^{(g)} - \bar{x}),
$$

-7-

$$
\lambda_1 = \left( \frac{3}{N} \prod_{g=1}^{3} \left| \frac{1}{N} v_{22.0, 1}^{(g)} \right|^{N/2} \right) \left( \frac{\left| \frac{1}{N} (I_{01} + I_{01}^{T}) (W + A) (I_{01} + I_{01}^{T})' \right|^2}{\frac{1}{N} (W + A)} \right)^{N/2},
$$

-7-

$$
\lambda_1 = \left( \frac{3}{N} \prod_{g=1}^{3} \left| \frac{1}{N} v_{22.0, 1}^{(g)} \right|^{N/2} \right) \left( \frac{\left| \frac{1}{N} (I_{01} + I_{01}^{T}) (W + A) (I_{01} + I_{01}^{T})' \right|^2}{\frac{1}{N} (W + A)} \right)^{N/2},
$$

-7-
\[ v_{20,1}^{(g)} = v_{22}^{(g)} - (v_{20}^{(g)}, v_{21}^{(g)}) \begin{pmatrix} v_{00}^{(g)} & v_{01}^{(g)} \\ v_{10}^{(g)} & v_{11}^{(g)} \end{pmatrix}^{-1} \begin{pmatrix} v_{00}^{(g)} \\ v_{10}^{(g)} \end{pmatrix}, \]

Note that

\[ \begin{pmatrix} v_{00}^{(g)} & v_{01}^{(g)} \\ v_{10}^{(g)} & v_{11}^{(g)} \end{pmatrix} = (I_{w^0 + r_1}, 0) v^{(g)}(I_{w^0 + r_0 + r_1}, 0)' \]

so that

\[ |\frac{1}{N} v_{22}^{(g)}| = |\frac{1}{N} v_{20,1}^{(g)}| |\frac{1}{N} (I_{w^0 + r_1}, 0) v^{(g)}(I_{w^0 + r_0 + r_1}, 0)'| \]

Thus, we may rewrite (2.5) in the form

\[ (2.7) \quad \lambda_1 = \frac{|\frac{1}{N} (I_{w^0 + r_1}, 0)(w^0 A)(I_{w^0 + r_0 + r_1}, 0)'|^N/2 \prod_{g=1}^{3} |\frac{1}{N} v_{22}^{(g)}|^N/2}{|\frac{1}{N} (w^0 A)|^N/2 \prod_{g=1}^{3} |\frac{1}{N} (I_{w^0 + r_1}, 0) v^{(g)}(I_{w^0 + r_0 + r_1}, 0)'|^N/2} \]

2.2. The likelihood ratio test statistic \( \lambda_2 \).

Comparing (2.4) to (1.5), we see that both the hypothesis \( H^* \) and the alternative hypothesis \( A \) place restrictions only on the parameters of the marginal distributions of the test score subvectors \( (x_0^{(g)}, x_1^{(g)}) \),
\( g = 1,2,3 \). Under these conditions, it is straightforwardly shown that

the LRT's \( \lambda_2 \) for testing \( H^* \) versus \( A \) can be found by reducing con-

sideration to that part of the sufficient statistic

\[
-\left( x_{0i}^g, x_{1i}^g, v_{00}^g, v_{01}^g, v_{11}^g \right)
\]

formed from the test score subvectors \( (x_{0i}^g, x_{1i}^g), i = 1,2,\ldots,N; g = 1,2,3 \).

That is, to find \( \lambda_2 \) we can act as if only \( (x_{0i}^g, x_{1i}^g) \) are observed,

\( i = 1,2,\ldots,N; g = 1,2,3 \). Under that assumption, we see from (1.5) and

(2.4) that \( (x_{0i}^1, x_{1i}^1) \) and \( (x_{0i}^2, x_{1i}^2) \) have identical distributions under

\( H^* \) and \( A \), and thus for testing \( H^* \) versus \( A \), groups 1 and 2 can be

combined into one group without loss of statistical generality. The testing

problem now becomes one of determining whether \( (T_0, U_1) \) and \( (T_0, U_3) \) are

parallel forms of the same psychological test, where scores on \( (T_0, U_1) \) are

obtained from \( N_1 + N_2 \) individuals, and scores on \( (T_0, U_3) \) are obtained

from \( N_3 \) individuals. From G-O, the LRTS for this problem (which, in

the notation of G-O, is

\( \lambda_{mv,c,v,m',c'} \) ) is

\[
(2.8) \quad \lambda_2 = \frac{|\frac{1}{N}(I_{r_0 r_0}'(W+B)(I_{r_0 r_0}^{-1})'\frac{N/2}{(N_1+N_2)(V(1)+V(2))_{11,0}}|^{2/2} \frac{1}{N_2} V(3)_{11,0}|^{2/2}}{\left|\frac{1}{N}(I_{r_0 r_1}'(W+B)(I_{r_0 r_1}^{-1})'\right|^{N/2}}
\]

where

\[
(2.9) \quad B = (N_1+N_2)(\frac{N_1}{N_1+N_2} - \bar{x}_1 + \frac{N_2}{N_1+N_2} \bar{x}_2 - \bar{x}) \cdot (\frac{N_1}{N_1+N_2} \bar{x}_1 + \frac{N_2}{N_1+N_2} \bar{x}_2 - \bar{x})' \cdot
\]

\[
+ N_3 (\bar{x}_3 - \bar{x}) (\bar{x}_3 - \bar{x})'
\]

\[-9-\]
and where for any matrix $C$,

$$
(2.10) \quad C = \begin{pmatrix}
  c_{00} & c_{01} & c_{02} \\
  c_{10} & c_{11} & c_{12} \\
  c_{20} & c_{21} & c_{22}
\end{pmatrix}
$$

blocked in the manner of $V^{(1)}, V^{(2)}$, etc., we have

$$
(2.11) \quad c_{11,0} = c_{11} - c_{10}c_{00}^{-1}c_{01}
$$

Recall that for $C$ as in (2.10),

$$
(2.12) \quad |c_{11,0}| = \frac{|\begin{pmatrix}
  c_{00} & c_{01} \\
  c_{10} & c_{11}
\end{pmatrix}|}{|c_{00}|} = \frac{|(I_{r_0+r_1,0})(I_{r_0+r_1,0})'|}{|c_{00}|}.
$$

Using this fact, we can modify (2.8) into a form more suitable for computation:

$$
(2.13) \quad \lambda_2 = \frac{\frac{1}{N} (I_{r_0,0})(W+B)(I_{r_0,0})'^N/2}{\frac{1}{N^2} (I_{r_0+r_1,0})(W+B)(I_{r_0+r_1,0})'^N/2} \frac{\frac{1}{N} (I_{r_0+r_1,0})(V^{(3)})(I_{r_0+r_1,0})'^N/2}{N/3 \cdot N/3} \frac{1}{N^2} \frac{1}{N^2} \cdot \frac{N/2}{N/2}.
$$

$$
\times \frac{\frac{1}{N^2} (I_{r_0+r_1,0})(V^{(1)})(I_{r_0+r_1,0})'^{(N^2+N^2)/2}}{\frac{1}{N^2} (V^{(1)})(V^{(2)})^N/2}.
$$
2.3. Calculation of the LRTS $\lambda$.

A comparison of (2.7) and (2.15) reveals that in calculating $\lambda = \lambda_1 \lambda_2$, there is little cancellation between terms in $\lambda_1$ and terms in $\lambda_2$. Thus, one reasonable way to compute $\lambda$ is to first compute $\lambda_1$ and $\lambda_2$ separately, and then multiply $\lambda_1$ and $\lambda_2$ to obtain $\lambda$. However, to take advantage of the one cancellation that does occur between terms in $\lambda_1$ and $\lambda_2$, we recommend first calculating $\lambda_1 u$ and $(\lambda_2 / u)$, where

$$u = \left| \frac{1}{N_3} \left( I_{r_0 + r_1} - I_{r_0} \right)^{N_3 / 2} \right| ;$$

the LRTS $\lambda$ can then be computed as the product of $\lambda_1 u$ and $(\lambda_2 / u)$.

In setting up the matrices $V(1), V(2), V(3), W, A, V(1)^{-1} A V(2)$, and $B$ for computation of $\lambda_1$ and $\lambda_2$, it is worth noting that some computational effort can be saved by taking advantage of the relationship

$$A = B + (N_1 + N_2)^{-1} N_1 N_2 \left( (1)_{x(1)} - x(2) \right); \quad (x(1) - x(2))$$

holding between $A$ and $B$ (compare (2.6) and (2.9)).

The rejection region for the null hypothesis $H$ based upon the test statistic $\lambda$ is of the form $\lambda < \lambda^*$, where $\lambda^*$ is a constant chosen so that the test of $H$ versus $A$ based on $\lambda$ has the desired level of significance $\alpha$. If we use the asymptotic approximation $-2 \log \lambda \sim \chi^2_f$, the degrees of freedom, $f$, for the approximation is equal to:

$$f = (r_0^2 + 2r_1 + r_2) + \frac{r_0 (r_0 + 1)}{2} + 2r_0 r_1 + 3r_0 r_2 + \frac{2r_1 (r_1 + 1)}{2} + 3r_1 r_2 + \frac{3r_2 (r_2 + 1)}{2} - r - \frac{r (r + 1)}{2}$$

$$= \frac{3}{2} r_1 + 3r_2 + r_0 r_1 + 2r_0 r_2 + 2r_1 r_2 + \frac{1}{2} r_1^2 + r_2^2 .$$

(2.14)
2.4. A Bartlett Modification.

In a somewhat different hypothesis testing context (that of testing homogeneity of variances), Bartlett (1937) suggested that the small sample behavior of the LRT might be improved by assuming in the calculation of the LRTS that the number of observations taken in each population equals the degrees of freedom left after estimating the various nuisance parameters. This idea was applied to broader testing contexts by Anderson (1958), and more recently by Gleser and Olkin (1972). Following the arguments in G-O, the Bartlett modification of $\lambda_1$ would replace $N_g$ by $N_g - r_0 - r_1 - l$ wherever $N_g$ explicitly appears in (2.7), $g = 1, 2, 3$. The Bartlett modification of $\lambda_2$ would replace $N_1 + N_2$ by $N_1 + N_2 - r_0 - l$ and $N_3$ by $N_3 - r_0 - l$ wherever $N_1 + N_2$ and $N_3$ explicitly appear in (2.13). To find the appropriate modification of $\lambda = \lambda_1 \lambda_2$, we follow a rule implicitly used by Anderson (1958) and G-O, and replace $N_g$ in the formula for $\lambda$ by the smaller of the substituted sample sizes (degrees of freedom) used in the modifications of $\lambda_1$ and $\lambda_2$, $g = 1, 2, 3$. In this particular problem, this rule means that $N_g$ is replaced by $N_g - r_0 - r_1 - l$ wherever $N_g$ explicitly appears in the formulas for $\lambda_1$ and $\lambda_2$, $g = 1, 2, 3$. Under this modification, $\lambda_1$ is replaced by

$$L_1 = \frac{\left|\frac{1}{m} \left( I_{1,0} + r_{1,0} \right) (W + A) (I_{1,0} + r_{1,0})'\right|^{\frac{m}{2}} \prod_{g=1}^{3} \left|\frac{1}{m} v(g)\right|^{\frac{m_g}{2}}}{\left|\frac{1}{m} (W + A)\right|^{\frac{m}{2}} \prod_{g=1}^{3} \left|\frac{1}{m} (I_{1,0} + r_{1,0}) v(g) (I_{1,0} + r_{1,0})'\right|^{\frac{m_g}{2}}}$$

where $m_g = N_g - r_0 - r_1 - l$, $g = 1, 2, 3$, and
\[ m = m_1 m_2 m_3 = N - 3(r_0 + r_1 + 1) \]

Similarly \( \lambda_2 \) is replaced by

\[
(2.16) \quad L_2 = \frac{\begin{vmatrix} \frac{1}{m} (I_{r_0 + r_1, o}) (W+B) (I_{r_0 + r_1, o})' \end{vmatrix}^2 \frac{1}{m_3} (V^{(3)}) (I_{r_0 + r_1, o})' \frac{m_3}{2}}{\begin{vmatrix} \frac{1}{m} (I_{r_0 + r_1, o}) (W+B) (I_{r_0 + r_1, o})' \end{vmatrix}^2 \frac{1}{m_3} V^{(3)} \frac{m_3}{2}}
\]

\[
\frac{1}{m_1 + m_2} (V^{(1)} + V^{(2)}) (I_{r_0 + r_1, o})' \left( \frac{m_1 + m_2}{2} \right)
\times \frac{1}{m_1 + m_2} (V^{(1)} + V^{(2)})(V^{(1)} + V^{(2)}) \left( \frac{m_1 + m_2}{2} \right)
\]

and the Bartlett modification of \( \lambda \) is

\[
(2.17) \quad L = L_1 L_2 = (L_1 u^*)(L_2 u^*)
\]

where

\[
u^* = \frac{1}{m_3} (I_{r_0 + r_1, o}) V^{(3)} (I_{r_0 + r_1, o})' \frac{m_3}{2}
\]

### 2.5. The rejection region for \( H \) based on \( L \)

The appropriate rejection region for the null hypothesis \( H \) based upon the test statistic \( L \) is of the form \( L < L^* \), where \( L^* \) is a constant chosen so that the test of \( H \) has a desired level of significance \( \alpha \).

The form of this rejection region parallels the form of the rejection region for the test of \( H \) versus \( A \) based on the LRTS \( \lambda \). Indeed, when \( N_1 = N_2 = N_3 = K \),
\[
\text{then } L = \frac{(K-r_0-r_1-1)}{K}, \text{ and thus when } L^* \text{ is set equal to } (\lambda^*)^{(K-r_0-r_1-1)/K},
\]
\[\lambda \text{ and } L \text{ define equivalent } \alpha \text{-level tests of } H. \text{ On the other hand, when}
\]
\[N_1, N_2, N_3 \text{ are not all equal to one another, } L \text{ and } \lambda \text{ are not monotonically related, and thus do not define equivalent } \alpha \text{-level tests of } H.
\]

To carry out the test of \(H\) versus \(A\) based on \(L\) it is necessary to find the value of the cut-off point \(L^*\) that will provide the desired level of significance \(\alpha\). Unfortunately, the distributional computations necessary in small samples to determine \(L^*\) are extremely complicated, and the results are still in an incomplete state. A similar comment applies to computation of the cut-off point \(\lambda^*\) that makes the test with rejection region \(\lambda < \lambda^*\) have level \(\alpha\) in small samples.

When \(N_1, N_2, N_3\) are all of reasonably large magnitude, a large-sample approximation can be used to find \(L^*\), namely

\[
(2.18) \quad L^* = \exp(-\frac{1}{2} \chi^2_f(\alpha)),
\]

where \(\chi^2_f(\alpha)\) is the \((1-\alpha)\)th fractile \((100(1-\alpha) \text{th percentile})\) of the \(\chi^2_f\) distribution, and \(f\) is given by \((2.14)\). Note that this large-sample approximation is identical to the large-sample approximation

\[
(2.19) \quad \lambda^* = \exp(-\frac{1}{2} \chi^2_f(\alpha))
\]

for \(\lambda^*\) (see Section 2.3). This, of course, is not surprising since \(L\) and \(\lambda\) have the same limiting distribution as \(N_g \to \infty, g = 1, 2, 3\). It should be noted, however, that in moderate samples the rejection regions \(\lambda < c\) and \(L < c\) are not the same, regardless of the constant \(c, 0 < c < 1\). For example, when \(N_1 = N_2 = N_3 = K\), then \(L = \lambda^{(K-r_0-r_1-1)/K}\), so that unless \((K-r_0-r_1-1)/K\) is close enough to 1, it is possible for the tests with the respective rejection regions \(\lambda < c\) and \(L < c\) to have a positive probability (under \(H\) and \(A\)) of reaching different conclusions.
3. An Illustrative Example

For the purposes of demonstrating application of the test statistic \( \lambda \) obtained in Section 2, data was obtained from 300 randomly selected answer sheets taken from the April, 1971 administration of the Scholastic Aptitude Test. By selecting various sections from the Scholastic Aptitude Test (SAT), three different test forms (actually test sub-forms, since not all sections were used) were constructed. Each form was equally represented (100 observations) in the sample. All constructed test forms had a common verbal section (subtest \( T_0 \)). Two of the forms had a common mathematics section (subtest \( U_1 \)), while the third form used a different mathematics section of the same SAT (subtest \( U_3 \)). The equating items from the SAT were combined into a third section which we can call the "equating section", which was assumed to differ among the three forms (thus creating subtests \( V_1, V_2, V_3 \)). Each section \( T, U, V \) on a given constructed form (subform) was summarized by a single score: \( T \) was summarized by \( x_0 \), \( U \) by \( x_1 \), and \( V \) by \( x_2 \). If the three forms (subforms) of the SAT are parallel, the parameters of the joint distributions of these subtest scores on each of the three forms must satisfy the null hypothesis \( H \). To test this null hypothesis against the alternative \( A \) of non-parallelism, we use the LRTS \( \lambda \) (Section 2) and reject \( H \) when

\[
\lambda < \exp\left(-\frac{1}{2} \chi_f^2(\alpha)\right),
\]

or equivalently when

\[
(3.1) \quad -2 \log \lambda > \chi_f^2(\alpha),
\]

where \( \alpha \) is the desired level of significance, \( f \) is given by (2.14), and \( \chi_f^2(\alpha) \) is the \((1-\alpha)\)th fractile of the \( \chi_f^2 \) distribution.

In the context of the given problem, \( r_0 = r_1 = r_2 = 1, r = 3, N_1 = N_2 \)
= N_3 = 100, and N = 300. The values of the sufficient statistic (\(x, \bar{V}\)) are given in Table 1.

Table 1. Summary of the Test Score Data

\[
\begin{align*}
\bar{x}^{(1)} & = (14.44, 12.64, 14.66), \\
\bar{x}^{(2)} & = (13.78, 12.86, 14.54), \\
\bar{x}^{(3)} & = (14.41, 10.36, 14.74), \\

V^{(1)} & = \begin{pmatrix}
6622.64 & 5260.84 & 5992.96 \\
5260.84 & 7525.04 & 5275.76 \\
5992.96 & 5275.76 & 7126.44
\end{pmatrix}, \\
V^{(2)} & = \begin{pmatrix}
4563.16 & 2965.92 & 4383.88 \\
2965.92 & 5248.04 & 2985.56 \\
4383.88 & 2985.56 & 5746.84
\end{pmatrix}, \\
V^{(3)} & = \begin{pmatrix}
6086.19 & 3055.24 & 5199.66 \\
3055.24 & 3545.04 & 2940.36 \\
5199.66 & 2940.36 & 6073.24
\end{pmatrix}.
\end{align*}
\]

From the data in Table 1, we find that

\[(3.2) \quad -2 \log \lambda = 34.02.\]

Since \(r_0 = r_1 = r_2 = 1\), we find from Equation (2.14) that \(f = 11\). Since \(\chi^2_{11}(0.005) = 26.8\), it follows from (3.1) and (3.2) that the null hypothesis of parallelism of the three constructed SAT subforms is rejected at the 0.5% level of significance.

The rejection of the parallelism hypothesis \(H\) in this particular example occurs largely because of the difference between the average scores of the examinees on the mathematics subtests \(U_1\) and \(U_3\) (see Table 1). Since in our construction of these subtests we used two different mathematics sections of the same SAT form, and since these two sections
are not supposed to be interchangeable (they presumably are designed to
test different abilities or levels of ability), the obtained differences
are not surprising. The rejection of the parallelism hypothesis thus
merely reflects the artificial way in which the three subforms in our
eexample were constructed, and should not be taken as an indication of any
lack of parallelism of the forms actually used in administering the SAT.

4. Generalizations

The techniques of Section 2 can straightforwardly be extended to
treat hypothesis testing problems in which scores on G forms of a test
are compared to one another in an attempt to determine if the G forms are
parallel. To present the relevant likelihood ratio test theory, however,
we need to change our notation, in order that the results may be given
in a compact form.

We assume that each test form consists of G subtests: \( S(1), S(2), \ldots, S(G) \). Subtest \( S(i) \) has \( i \) different versions: \( S_1(i), S_{G-i+2}(i), \ldots, S_{G-1}(i), S_G(i) \),
where \( S_1(i) \) is common to test forms \( 1, \ldots, G-i+1 \), and version \( S_j(i) \)
appears in test form \( j, j = G-i+2, \ldots, G \). Thus, subtest \( S(1) \) has only one
version \( S_1(1) \) which appears in all test forms. Subtest \( S(2) \) has 2 versions
\( S_1(2) \) and \( S_G(2) \), with \( S_1(2) \) appearing in test forms \( 1, 2, \ldots, G-1 \) and \( S_G(2) \)
appearing only in test form \( G \). Finally, \( S(G) \) has \( G \) different versions
\( S_1(G), S_2(G), \ldots, S_G(G) \), each version appearing in one and only one test
form. The assignment of subtest versions to test forms is illustrated
in Figure 1.
\begin{tabular}{|c|c|c|c|c|c|c|}
\hline
 & Subtest 1 & Subtest 2 & Subtest 3 & $\ldots$ & Subtest G-1 & Subtest G \\
\hline
Test Form 1 & $S_1(1)$ & $S_1(2)$ & $S_1(3)$ & $\ldots$ & $S_1(G-1)$ & $S_1(G)$ \\
\hline
Test Form 2 & $S_1(1)$ & $S_1(2)$ & $S_1(3)$ & $\ldots$ & $S_1(G-1)$ & $S_2(G)$ \\
\hline
Test Form 3 & $S_1(1)$ & $S_1(2)$ & $S_1(3)$ & $\ldots$ & $S_3(G-1)$ & $S_3(G)$ \\
\hline
$\vdots$ & $\vdots$ & $\vdots$ & $\vdots$ & $\vdots$ & $\vdots$ & $\vdots$ \\
\hline
Test Form G-1 & $S_1(1)$ & $S_1(2)$ & $S_{G-1}(3)$ & $\ldots$ & $S_{G-1}(G-1)$ & $S_{G-1}(G)$ \\
\hline
Test Form G & $S_1(1)$ & $S_G(2)$ & $S_G(3)$ & $\ldots$ & $S_G(G-1)$ & $S_G(G)$ \\
\hline
\end{tabular}

Figure 1. Hierarchical assignment of subtest forms to the $G$ test forms.

Let each of the $G$ test forms be characterized by $r = \sum_{g=1}^{G} r_g$ scores:

$r_1$ on subtest $S(1)$, $r_2$ on subtest $S(2)$, $\ldots$, $r_G$ on subtest $S(G)$. Let $x_1^{(g)}$, $x_2^{(g)}$, $\ldots$, $x_G^{(g)}$ be the scores of a typical individual on subtests $S(1)$, $S(2)$, $\ldots$, $S(G)$, respectively, of test form $g$. Hence $x_i^{(g)}$ is an $1 \times r_i$ vector, $i = 1, 2, \ldots, G$. Let

\begin{equation}
(x^{(g)}) = (x_1^{(g)}, x_2^{(g)}, \ldots, x_G^{(g)})
\end{equation}

be the $1 \times r$ vector of test scores by a typical individual who takes test form $g$. As before, we assume that $x^{(g)}$ has a multivariate normal distribution with mean vector

\begin{equation}
\mu^{(g)} = (\mu_1^{(g)}, \mu_2^{(g)}, \ldots, \mu_G^{(g)})
\end{equation}

and covariance matrix
\begin{equation}
\Sigma^{(g)} = \begin{pmatrix}
\Sigma_{11}^{(g)} & \Sigma_{12}^{(g)} & \ldots & \Sigma_{1G}^{(g)} \\
\Sigma_{21}^{(g)} & \Sigma_{22}^{(g)} & \ldots & \Sigma_{2G}^{(g)} \\
\vdots & \vdots & \ddots & \vdots \\
\Sigma_{G1}^{(g)} & \Sigma_{G2}^{(g)} & \ldots & \Sigma_{GG}^{(g)}
\end{pmatrix},
\end{equation}

where the blocking of \( x^{(g)} \) and \( \Sigma^{(g)} \) conforms to the blocking of \( x^{(g)} \).

That is, \( \mu_{j}^{(g)} \) is \( 1 \times r_{j} \) and \( \Sigma_{jk}^{(g)} \) is \( r_{j} \times r_{k} \); \( j, k = 1, 2, \ldots, G \).

We assume that individuals have been assigned to test forms at random in such a way that \( N_{g} \) individuals take test form \( g \), \( g = 1, 2, \ldots, G \).

We also assume that the conditions under which individuals are examined are identical, and that individuals work independently of one another.

Let the score vector of the \( i \)th individual taking test form \( g \) be

\begin{equation}
x_{i}^{(g)} = (x_{1i}^{(g)}, x_{2i}^{(g)}, \ldots, x_{Gi}^{(g)}).
\end{equation}

Under our above-stated assumptions the \( x_{i}^{(g)} \), \( i = 1, 2, \ldots, N_{g} \); \( g = 1, 2, \ldots, G \), are mutually statistically independent, and a sufficient statistic for the parameters of the distributions of test scores on the \( G \) test forms is \( \overline{x}, V \), where \( \overline{x} = (\overline{x}^{(1)}, \overline{x}^{(2)}, \ldots, \overline{x}^{(G)}) \), \( V = (V^{(1)}, V^{(2)}, \ldots, V^{(G)}) \),

\begin{equation}
\overline{x}^{(g)} = \frac{1}{N_{g}} \sum_{i=1}^{N_{g}} x_{i}^{(g)} = (\overline{x}_{1}^{(g)}, \overline{x}_{2}^{(g)}, \ldots, \overline{x}_{G}^{(g)}),
\end{equation}

and

\begin{equation}
V^{(g)} = \frac{1}{G} \sum_{i=1}^{N} (x_{i}^{(g)} - \overline{x}^{(g)})(x_{i}^{(g)} - \overline{x}^{(g)}),
\end{equation}

\begin{equation}
\begin{pmatrix}
V_{11}^{(g)} & V_{12}^{(g)} & \ldots & V_{1G}^{(g)} \\
V_{21}^{(g)} & V_{22}^{(g)} & \ldots & V_{2G}^{(g)} \\
\vdots & \vdots & \ddots & \vdots \\
V_{G1}^{(g)} & V_{G2}^{(g)} & \ldots & V_{GG}^{(g)}
\end{pmatrix},
\end{equation}

\( g = 1, 2, \ldots, G \).
If the G test forms are not parallel, the construction of our experimental design assures us that at least the following relationships hold among the parameters:

\[
\begin{align*}
A: & \quad \mu_j^{(1)} = \mu_j^{(2)} = \ldots = \mu_j^{(G-j+1)}, \quad j = 1, 2, \ldots, G, \\
& \quad \Sigma_j^{(1)} = \Sigma_j^{(2)} = \ldots = \Sigma_j^{(G-j+1)}, \quad k \leq j, \quad j = 1, 2, \ldots, G.
\end{align*}
\]

(4.7)

On the other hand, if the G forms are parallel, then the following hypothesis about the parameters holds:

\[
H: \quad \mu^{(1)} = \mu^{(2)} = \ldots = \mu^{(G)}, \quad \Sigma^{(1)} = \Sigma^{(2)} = \ldots = \Sigma^{(G)}.
\]

(4.8)

The likelihood ratio test for testing the null hypothesis H against the alternative A is constructed in terms of \(\bar{x}^{(1)}, \bar{x}^{(2)}, \ldots, \bar{x}^{(G)}, \mathcal{V}(1), \mathcal{V}(2), \ldots, \mathcal{V}(G)\), and the following quantities:

(4.9) \[ M_j = \sum_{g=1}^{j} N_g, \quad j = 1, 2, \ldots, G, \]

(4.10) \[ q_j = \sum_{g=1}^{j} r_g, \quad j = 1, 2, \ldots, G, \]

(4.11) \[ E_j = (I_{q_j}, 0): q_j \times r, \quad j = 1, 2, \ldots, G, \]

(4.12) \[ W_j = \sum_{g=1}^{j} \mathcal{V}(g), \quad j = 1, 2, \ldots, G, \]

and \( B_1 = 0 \),

(4.13) \[ B_j = B_{j-1} + \frac{N_j M_{j-1}}{M_j} \left( \frac{1}{M_j} \sum_{g=1}^{j-1} N_g \bar{x}(g) - \bar{x}(j) \right) \left( \frac{1}{M_j} \sum_{g=1}^{j-1} N_g \bar{x}(g) - \bar{x}(j) \right), \]

for \( j = 2, 3, \ldots, G \). Note that

(4.14) \[ r = q_G = \sum_{g=1}^{G} r_g, \quad N = M_G = \sum_{g=1}^{G} N_g. \]
The likelihood ratio test statistic \( \lambda \) for testing \( H \) versus \( A \) can now be derived by a simple extension of the method used in Section 2 (for the case \( G = 3 \)). The resulting LRTS is

\[
(4.15) \quad \lambda = \left( \frac{\prod_{j=2}^{G} \left| \frac{1}{M_j} E_{j-1}^\prime \left( W_{G} B_j \right) E_{j-1}^\prime \right|^{M_j/2} \frac{1}{N_j} V(j)^{N_j/2}}{\prod_{j=2}^{G-1} \left| \frac{1}{M_j} E_{j-1}^\prime \left( W_{G-j+1} B_j \right) E_{j-1}^\prime \right|^{M_j/2} \frac{1}{N_j} E_{G-j+1}^\prime V(j) E_{G-j+1}^\prime \right)^{N_1/2} \]

\[
\times \left( \frac{\prod_{j=2}^{G-1} \left| \frac{1}{M_j} E_{j-1}^\prime \left( W_{G-j+1} B_j \right) E_{j-1}^\prime \right|^{M_j/2} \frac{1}{N_j} V(j)^{N_j/2}}{\prod_{j=2}^{G-1} \left| \frac{1}{M_j} E_{j-1}^\prime \left( W_{G-j+1} B_j \right) E_{j-1}^\prime \right|^{M_j/2} \frac{1}{N_j} E_{G-j+1}^\prime V(j) E_{G-j+1}^\prime \right)^{N_1/2} \right)
\]

We reject the null hypothesis \( H \) if

\[
(4.16) \quad \lambda < \lambda^*
\]

where \( \lambda^* \) is chosen so as to give the test a desired level of significance \( \alpha \). For \( N_1, N_2, \ldots, N_G \) all moderately large, we can use the fact that under \( H \)

\[
(4.17) \quad -2 \log \lambda \xrightarrow{\text{law}} \chi^2_f,
\]

where

\[
(4.18) \quad f = \sum_{g=2}^{G} \frac{r_g (r_g + r_{g-1} + 3)}{2} + \frac{r_1 (r_1 + 3)}{2} - \frac{1}{2} \left( \sum_{g=1}^{G} r_g \right) \left( \sum_{g=1}^{G} r_g + 3 \right),
\]

to find an approximate level-\( \alpha \) test based on \( \lambda \). This test rejects \( H \) when

\[
(4.19) \quad -2 \log \lambda > \chi^2_f(\alpha),
\]

where \( \lambda \) is given by (4.15), and \( f \) is given by (4.18).
Various other hierarchical designs for testing the parallelism psychological tests can be analyzed using the methods of likelihood ratio testing developed here and in the earlier paper (Gleser and Olkin (1972)). Discussion of these designs and their analysis is planned for a future paper.

ACKNOWLEDGEMENTS

This work was supported in part by the Educational Testing Service, The Air Force Office of Scientific Research (contract F44620-70-C-0066 at Johns Hopkins University and grant AFOSR 73-2432 at Purdue University) and the National Science Foundation (Grant GP-32326 X) at Stanford University. We would like to express our thanks to William H. Angoff and June Stern for their help in obtaining the data of Table 1, and to Britt-Marie Alsén for her help in performing the computations.

REFERENCES

