Advances in Methodology, Data Analysis, and Statistics
Anuška Ferligoj (Editor), Metodološki zvezki, 14
Ljubljana: FDV, 1998

# An Application of Factorial Invariance in the Context of a Step-down MANOVA with Latent Variables

Brendan P. Bunting[1] and Eugene Mooney[2]

**Abstract**

In MANOVA two different variable systems are possible, emergent and latent. This paper sets out an analysis using a latent variable approach in the context of a study which examines the effects of coaching and practice on test performance. Seven parallel forms of a test were administered to school children (aged 10-11 years) under two conditions. In the first condition (group one) children were given three hours of coaching after having taken three selection tests (n = 241). The second group received coaching prior to the administration of the tests (n = 311). Five of the seven tests were administered over a two week period and the remaining two were given some nine months later. The analysis was conducted with latent variables in a step-down MANOVA model. This permitted an examination of the psychometric properties of the measures through the testing of assumptions which are usually left unstated when an emergent variable system is used in conjunction with the MANOVA model. Coaching and practice are shown to have substantial effects. However, the interpretation of the long-term effects is confounded because of unequal intercepts. The identification of this differential bias is seen, among others results, as being a positive advantage of the latent variable approach to the analysis of mean structured data.

## 1 Introduction

This paper details the statistical analyses used to examine the effects of coaching and practice over two experimental conditions, and during a nine months period. Seven measures were used in the analysis. Three measures are employed in the first experimental stage, where one group received three hours coaching and the other did

[1] Psychology, University of Ulster Jordanstown, Newtownabbey, Co Antrim BT37 0QB
[2] Northern Ireland Statistics & Research Agency, Parliament Buildings, Stormont Estate, Belfast BT4 3SW

not. In the second stage two measures were used. By this second experimental stage both groups had received the benefits of coaching. In a third and final stage, some nine months later, two further tests were administered. Within this analysis our interest was in (a) the psychometric properties of the measures and (b) the extent of change occurring both within and between the two experimental conditions.

The use of multiple measures, such as those employed in all three stages of this experiment, can be seen as a variable system. Such a collection of measures has been described by Huberty and Morris (1989) as:

> "A system of outcome variables {which} may be loosely defined as a selection of conceptually interrelated variables that, at least partially, determines one or more meaningful underlying variates or constructs" (p.304)

There are many well known techniques for handling variable systems, e.g., factor analysis, principal components, discriminant analysis and canonical correlation. Indeed, in many multivariate situations the main goal of the analysis is to examine the behaviour of the underlying variable system. This is also the case in research where there are multiple dependent variables. The preferred model in such situations has often been Multivariate Analysis of Variance (MANOVA) with an emergent variable system. This choice is frequently justified in so far as the outcome measures are conceptually related and statistically correlated. In this event they contain redundant information, which within an univariate ANOVA context will lead to the duplication of results (Van de Geer, 1971, p. 271).

However, the most appropriate way to handle the variable system within MANOVA is open to debate. Traditionally, the underlying structure of the measures, within MANOVA, is assumed to be that of an emergent variable system (Cole, Maxwell, Arvey & Salas, 1993, Bollen and Lennox, 1991). The variables are therefore assumed to determine the construct(s). Such variable systems require that all relevant measures of the construct be included (Cook & Campbell, 1979, p.65; Cole et al., 1993; Bollen & Lennox, 1991). It is also assumed that the disturbance terms are uncorrelated. The presence of correlated disturbance terms, in the MANOVA model, has been shown by Cole et al. (1993) to produce misleading discriminant function coefficients. A further limitation with the emergent variable system, employed within the MANOVA, is that the procedure does not detect the presence of differential functioning across the observed measures. In other words, subjects from different groups could have the same true score on the construct but this may have been obtained from very different patterns of responding on the observed measures.

Huberty and Morris (1989) in an examination of the application of MANOVA to emergent variable systems, as they were employed in five American Journals, reported that from the 222 articles where multiple outcome measures were present

few authors reported any interest in the variable system underlying the mean comparisons. In part this may be because of the lack of control which the researcher has over the variable system(s) and the exploratory nature of the proposed structure for the underlying variable system(s).

A well-known alternative to the emergent variable system, employed within the MANOVA, is that of latent variables. In the latent variable approach, responses to the observed measures are dependent upon the latent construct. The decision of which variable system to use (emergent or latent) will largely depend upon the construct of interest (Bollen & Lennox 1991). Where the constructs are unifactorial and tau-equivalent across groups then both the MANOVA and SEM approach will yield essentially equivalent interpretations (Cole et al., 1993). However, Cole et al (1993) point to the unanswered question of power in the MANOVA and SEM models.

> "In any congeneric variable system, increased reliability and validity of the measures will be reflected in increased correlations between the measures. Increased intercorrelations can diminish the power of MANOVA (e.g., Ramsey, 1982). It is not obvious that SEM would be similarly affected" (p. 183).

In this analysis the measures within the variable systems are all highly intercorrelated.

Cole et al. (1993) have also pointed to other considerations which a researcher might also wish to take into consideration when making a choice between an emergent and latent variable system.

> "In summary, we speculate that MANOVA would be especially appropriate for data sets in which the phenomena under investigation were reflected in emergent variable systems *and the investigator was satisfied that larger mean differences between groups were reflective of true group differences (and that bias did not differentially affect some measures more than others).* On the other hand, we recommend SEM when the phenomena under investigation are reflected in latent variable systems, *when the investigators wish to examine the within group dependent variable structure, or the researchers want to test for differential measurement bias across groups*"(p. 175). (Italics added.)

Many of these known limitations, in the use of MANOVA with emergent variables, can be overcome through the use of structural equation modelling (SEM) (Bagozzi and Yi, 1989). Within the present analysis a latent variable system is thought most appropriate for a variety of reasons. Firstly, the outcome measures are achievement tests and the responses to such measures are generally seen as having arisen for an underlying construct (Spearman, 1904). Secondly, due to the nature of the experimental design considerable control is required over the variable systems.

Thirdly, the reliability of the measures is of considerable interest since it not only has direct consequences for the parameter estimates but also the selection of individuals (Bunting, Saris and McCormack, 1987). (The measures employed in this study were used for the process of educational selection.) Fourthly, it is not possible to rule out, a priori, the presence of correlated errors. These could arise from either the experimental treatment or some commonality between the tests within each condition. Fifthly, the potential biasing effects of test differential functioning across the experimental conditions needs to be taken into consideration (Rock, Werts and Flaugher, 1978).

# 2 Modelling invariance in the context of a step-down MANOVA

## 2.1 Data structure

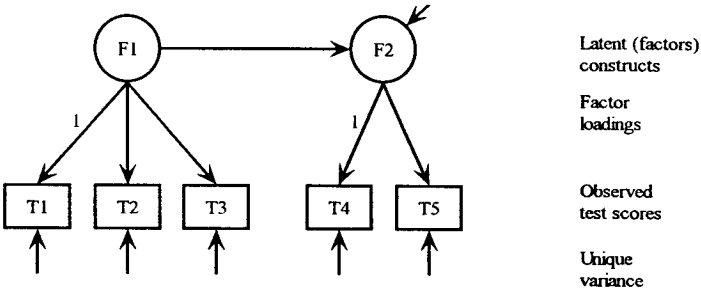
A series of five parallel forms of a verbal reasoning test was administered to ten-year old children, from a sample of schools in Northern Ireland, ten months prior to taking their selection tests to decide the type of secondary education they should obtain. The first five tests were administered two to three days apart, over a period of two weeks. The tests were presented in a Latin-square design in order to randomise the order of presentation. A further two tests (tests 6 and 7) were administered nine months later. These latter two papers were taken one week apart and were supplied by the Department of Education in Northern Ireland to all schools to be administered as part of the preparation for the actual selection procedure. Children therefore had no opportunity to see these tests before administration. Unlike the five measures administered nine months earlier, scores on tests six and seven reflected performance on only a single test.

Two conditions were used in the study. In condition one, coaching for a period of three hours occurred after the third test administration (group one). In the second condition, three hours coaching occurred before any of the five tests were administered (group two). The assignment to one or other condition was done at random.

## 2.2 Analysis

The analysis was conducted in three stages. In the first, the homogeneity of the variance-covariance matrices across the conditions was examined, using all seven measures in each group. At the second stage, the invariance of the first five tests was examined in terms of a two factor (latent variable) model in each experimental

condition. The first factor represented the first three measures, while the second factor had two observed measures.




F1 and F2: Represent the latent constructs (factors) at stages one and two of the experimental conditions.
T1 - T5: Summary scores for the measures at five points in time.
The lines linking the factors to the observed variables (summary scores) represent the factor loadings and are contained in the matrix lambda ($\Lambda$).
The line linking F1 to F2 represents the regression coefficient, contained in the beta matrix ($\beta$).

Figure 1: A diagrammatic representation of the model for the invariance of the first two factors (an identical structure was present in both groups).

In the third stage of the analysis all seven observed measures are employed within each condition. The model from stage two (Figure 1) will be extended to include a third factor created from the sixth and seventh observed measure (see Figure 2). The measures for this third factor were obtained some nine months later.

# 3 Results

### Stage 1. Homogeneity of the variance-covariance matrices across groups

(1) $\mathbf{H_{form}}$

In this model the var-covariance matrices for each group was constrained to be equal. This model did not adequately describe the data ($\chi^2$ = 88.14, df = 28, p = .000).

In general when multiple groups are compared it is assumed that the variances and covariances are equal across the groups. This is not a necessary assumption when MANOVA is modelled as a latent variable system (Kühnel, 1988).

**Stage 2. Tests of factorial and structural invariance in the two factor model**

The structure of this model was tested using a number of constraints across the groups.

(2) $H_{\Lambda}$ **Invariant factor loadings**

This is seen as the minimum condition which must hold before testing restrictions on means and intercepts (Bollen, 1989, p.366).

(3) $H_{\Lambda, \psi,}$ Invariant factor loadings **and factor variances**

Where this holds one can then infer the stability of scores both within and between groups.

(4) $H_{\Lambda, \psi, \beta}$ Invariant factor loadings, factor variances, **and structural coefficient**

The rate of change is then tested across the experimental conditions. This assumption (like many others) is frequently untested when MANOVA is used with emergent variable systems.

(5) $H_{\Lambda, \psi, \beta, \tau}$ Invariant factor loadings, factor variances, structural coefficient, **and intercepts**

By restricting the intercepts across the groups it will be possible to test for the presence of differential test (item) functioning across the groups

(6) $H_{\Lambda, \psi, \beta, \tau, \alpha}$ Invariant factor loadings, factor variances, structural coefficient, intercepts, **and means.**

In MANOVA designs with emergent variable systems the main focus of the analysis is on mean differences, ignoring the assumption of invariance implicit in the analysis. However, from a structural equation perspective it is only after the psychometric properties of the measures have been examined that the focus turns to an examination of the means.

**Stage 2. An invariant two factor model**

The chi-square value for the test of invariant factor loadings ($H_{\Lambda}$ ) was 29.257 with 11 df and p=.002 (RMSEA = .055: 90% CI, .0312; .0795). This was taken as a reasonable description of the data. However, the residual variance of the second factor in group 2, that is, those who had received coaching prior to taking any tests

was negative. The value for this residual variance was -.020 in the 'within group completely standardised solution'. The comparable value in the other group was .012. This suggests that there is little or no unexplained variance in factor 2 for both groups.

The residual variance in factor 2 for both groups was restricted to zero ($\psi_{22} = 0$). This model has 13 df and a chi-square value of 29.934, p=.005 (RMSEA = .049: 90% CI, .026; .072) and for the present purpose this model is accepted as an adequate description of the data. Since the factor loadings were restricted to be equal across the groups this is equivalent to a test of equality of scaling units. The highest modification index was 8.31 for $\theta_{4,2}$ in group one. It thus appears that a two factor model can be used to describe the relationship between the first five observed variables in both groups. Further, this result implies that the respondents' rankings, in both groups, have remained stable. Coaching does not appear to have an effect on the stability of responses to the test questions over time.

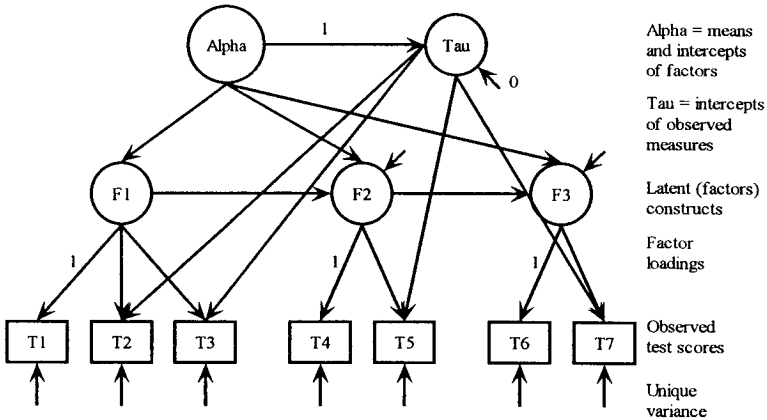
Restricting the structural coefficients ($\beta_{2,1}$) equal across the groups produced a chi-square of 30.340, df=14, p=.007 (RMSEA =.046: 90% CI, .023; .069). Of course, this does not tell us anything about possible mean changes brought about by coaching or other factors.

Before testing for the equality of factor means (factor one) and the intercepts for the second factor between groups 1 and 2 the intercepts of the observed measures were constrained to be equal across the groups. This restriction was made to check for possible test (observed measures conditioned on the factor(s)) differential functioning across the groups. The resulting model had a chi-square of 33.326, df=18, p=.015 (RMSEA = .039: 90% CI, .017; .06), which indicates that neither the factor means nor intercepts are likely to be confounded by the potential biasing effects of differential response patterns in the tests across the groups.

When the factor means in the first wave, and intercepts in the second wave, were restricted to be equal across the groups, the chi-square value substantially increased ($\chi^2 = 99.856$, df = 20, p = .000). A second test was then conducted where only the intercepts across the second factor were restricted to be equal. The chi-square for this model was 49.603, df=19, p=.000 (RMSEA = .054: 90% CI, .036; .073). This led to a substantial reduction in the chi-square value for the loss of one degree of freedom. The expected mean value, within each group, can be obtained for this second factor by multiplying the slope by the mean value for factor one and adding the intercept value. These values are shown in Table 2.

**Stage 3. An invariant three factor model**

In stage 3 of the analysis the results obtained from two tests (tests 6 and 7), administered some nine months later, were incorporated into the previous model as a new factor (Figure 2).




F1 - F3: Factors one to three represent the latent constructs in each of the experimental stages. The variances for these were tested in the matrix psi.

T1 - T7: The summary scores for the measures at seven points in time.

The lines linking the factors to the observed variables (summary scores) represent the factor loadings and are contained in the matrix lambda ($\Lambda$).

The line linking F1 to F2 and from F2 to F3 represents the regression coefficient, contained in the beta matrix ($\beta$).

Tau ($\tau$): This factor was generated to contain the information regarding the intercepts of the observed measures. For the purpose of identification the first intercept within each factor is not obtained.

Alpha ($\alpha$): This factor contains information for computing the factor means.

Figure 2: A diagrammatic representation of a step-down MANOVA design within the context of latent variables.

In the first of these models the factor loadings, variances and structural coefficients were set equal across the groups. This model was a reasonable representation of the data ($\chi^2$ = 60.919, df=34, p=.003 and RMSEA = .038: 90% CI, .022; .053).

The intercepts for the observed measures were then restricted to be equal across the groups. This model produced a chi-square value of 69.834, with 38 df and p=.001 (RMSEA = .039: 90% CI, .024; .053). In strict statistical terms this is a significantly less satisfactory description of the data than that provided by the previous two factor model and would suggest that the intercept(s) for the last two observed measures was not equal across the groups. Hence, an interpretation of factor means, at least for this third factor, is problematic (this problem is discussed by Cole, Maxwell, Arvey and Sala, 1993). When the intercept for the seventh measure is left

unrestricted, the chi-square value drops to 63.317 with 37 df, p=.005 (RMSEA = .036: 90% CI, .020; .051).

# 4 Reliabilities and predicted factor means

With one or two exceptions the reliabilities of the measures seem to be moderate. One such exception is the reliability of the final measure in group 2. It is difficult to account for this drop in reliability but it is noteworthy that this is one of the measures where the intercepts were unequal across the groups.

Table 2: Results from the final model for the group who received three hours coaching after the third test (group 1) are shown first, while the parameter estimates for the group who received coaching prior to taking any tests (group 2) are indicated second and shown in **bold**.

| Factor One | Factor two | Factor three | $Beta_{21}$ | $Beta_{32}$ | Psi (residuals) | Factor intercepts | Factor means | Reliability |
|---|---|---|---|---|---|---|---|---|
| Standardised solution (common metric) | | | | | Unstandardised values | | | |
| .826 | .873 | .932 | 1.152 | 1.221 | 91.881 | 24.881 | 24.881 | T1 .74 |
| | | | | | | **31.248** | **31.248** | **T1 .65** |
| .784 | .867 | .848 | | | ------- | 1.347 | 30.010 | T2 .55 |
| | | | | | | **0.909** | **35.089** | **T2 .68** |
| .873 | | | | | 72.505 | 17.516 | 54.158 | T3 .72 |
| | | | | | | **17.668** | **60.512** | **T3 .80** |
| | | | | | | | | T4 .75 |
| | | | | | | | | **T4 .77** |
| | | | | | | | | T5 .79 |
| | | | | | | | | **T5 .72** |
| | | | | | | | | T6 .85 |
| | | | | | | | | **T6 .88** |
| | | | | | | | | T7 .81 |
| | | | | | | | | **T7 .66** |

T1, T2,....,T7 represent the tests, one to seven

The intercept values for the first factors are equivalent to the means. Therefore, as shown in the table above, those who received coaching prior to taking any tests (groups 2) had a mean value of 31.248, while those without the benefits of coaching prior to taking the first three tests (groups 1) obtained a mean value of 24.881 on the first factor. The mean score on the second factor was 35.089 for group 2 and 30.010 for group 1. By time three the predicted scores were 60.512 for group 2 and 54.158

for group 1. However, given the presence of differential item functioning it was thought best to also conduct an analysis using only the seventh measure on the third factor. The result indicated that the gap in test scores between groups one and two had not only narrowed but was now in favour of those in group one - a lead of some four points. Such a result is obtained only if the seventh measure is used as an indicator of the third factor and the respective reliabilities are inserted. A similar analysis conducted with the sixth measure indicated a two point lead in favour of group two. The observed differences are shown in appendix one.

# 5 Discussion

The decision of whether data should be modelled as an latent variable system or as an emergent variable is one that should be based upon the nature of the construct under consideration. However, in some situations this may not be a simple either/or decision. In others, the researcher may be setting about the creation of measures for a construct. At present there would seem to be considerable advantages to using the latent variable approach. The present analysis has illustrated a number of advantages associated with the employment of latent variables. Foremost among these is the ability to assess the psychometric properties of the measures. In the present analysis the measures were shown to be fairly reliable.

The var-covariance matrix was not equal across the groups. At this point it would have been possible to explore the nature of this inequality but since it was not an essential condition for a latent variable approach, the analysis moved on to the next stage. In this second stage it was shown that the factor loadings were equivalent across the groups. This model held for the three measures of the construct and then again for the next two. It was then clear that the same metric was present under very different experimental conditions.

The rank order of individual was then assessed by restricting the factor variances to be equal across the groups. This indicated that although the pupils were subjected to very different experimental treatments their rank order remained stable. In the next stage of the analysis it was possible to determine the extent of change across the conditions. The initial difference between the groups on the first construct was over six points. After those in the second group had received three hours coaching this gap was only narrowed by under two points. A number of reasons can be postulated for the remaining differential across the groups; however, we will avoid this in the present context. What has been established is that the formal test of equality in the slopes does represent what is happening in both conditions.

In the final stage of the analysis, nine months later, the pupils took two test which were administered as part of the educational selection process. Again we were in a position to test the influence of coaching only this time not just over a period of three hours coaching but within a period of nine months. Teachers were encouraged to keep coaching to a minimum by officials from the department of education, but

this is almost completely ignored by teachers. So after this sustained period of coaching an evaluation of the means was again possible. This showed that both groups had dramatically improved their scores by around 25 to 31 points.

However, the extent of this change both within and across groups very much depended upon which observed measure was used to set the metric of the latent variable. The first of the Department's tests had a reliability of .85 in the group who received three hours coaching only after the first three tests (group 1), while those in the group where coaching occurred before the test had a reliability of .88. If the groups were to be compared using only this measure (and a correction made for the reliability) then the change from the second latent factor to the third, some nine months later, indicated a difference of over two points in favour of group 2 (those who received coaching prior to taking any tests).

A very different story is told if the sixth measure is removed and only the seventh is included within the analysis (again corrected for measurement error). The very different reliability shown for the two groups (.81 for group one and .66 for group two) produces a result which indicates that the experimental group which had done least well on the first two factors, had not only closed the gap, but were some four points ahead.

The results from factor one and two, both within and across conditions, are fairly unambiguous. The group who received coaching prior to the tests (group 2) had a considerable lead in test scores - some six points. When group one received three hours coaching the gap narrowed by nearly two points. It can be seen that the group who received coaching prior to taking any tests (group 2) have continued to gain in scores (by some 4-5 points), even though no further coaching has been given. Thus, the effects of coaching and practice, in combination, would appear to have added value. In part this may explain why the initial gap between the groups was not closed at stage two.

However, it is at stage three, some nine months later that considerable ambiguity arises in the results. It was formally established that the intercepts for this third factor were not equal across the groups. Hence, establishing mean differences is fraught with ambiguity. Very different results can be obtained depending upon the way the measures are examined. In the present analysis this problem is easily detected because the MANOVA analysis has been formulated as a latent variable model with a structural equation model. On the other hand, if the analysis had been conducted using an emergent variable system this problem would not only have remained undetected, but would have as noted by Cole et al. (1993):

> "If a particular measure discriminates between groups more effectively than would be expected given its relation to the underlying construct of interest, MANOVA simply increases its discriminant coefficient" (p. 183).

Note: This paper has greatly benefited from the comments and advice of Germà Coenders and an anonymous reviewer. My thanks also to Anuška Ferligoj for encouraging me to present this work.

# 6 References

[1] Bagozzi, R. P. and Yi, Y. (1989): On the use of structural equation models in experimental designs. *Journal of Marketing Research*, **26**, 271-284.

[2] Bunting, B. Saris, W. E. and McCormack, J. (1987): A second-order factor analysis of the reliability and validity of the 11-plus examination in Northern Ireland. *Economic and Social Review*, **18**, 137-147.

[3] Bollen, K. (1989): *Structural Equations with Latent Variables*. Wiley & Sons, Inc. New York.

[4] Bollen, K. and Lennox, R. (1991): Conventional wisdom on measurement: A structural equation perspective. *Psychological Bulletin*, **110**, 305-314.

[5] Cole, D. A., Maxwell, S. E., Arvey, R. and Salas, E. (1993): Multivariate group comparisons of variable systems: MANOVA and structural equation modeling. *Psychological Bulletin*, **114**, 174-184.

[6] Cook, T. D. and Campbell, D. T. (1979): *Quasi-experimentation: Design & Analysis Issues for Field Settings*. Chicago. Rand McNally.

[7] Huberty, C. J. and Morris, J. D. (1989): Multivariate analysis versus multiple analyses. *Psychological Bulletin*, **76**, 49-57.

[8] Kühnel, S. M. (1988): Testing MANOVA designs with LISREL. *Sociological Methods and Research*, **16**, 504-523.

[9] Rock, D. A. Werts, C. E. and Flaugher, R. L. (1978): The use of analysis of covariance structures for comparing the psychometric properties of multiple variables across populations, *Multivariate Behavioral Research*, **13**, 403-418.

[10] Spearman, C. (1904): General intelligence objectively determined and measured. *American Journal of Psychology*, **15**, 201-293.

[12] Van de Geer, J. P. (1971): *Introduction to Multivariate Analysis for the Social*. W. H. Freeman and Company, San Francisco.

# Appendix 1

Correlations, standard deviations and means for both those who received coaching after the third test (below the diagonal, group 1) and for those where coaching occurs prior to the tests (above the diagonal, group 2).

Correlations
1.0 .6743 .7020 .7252 .6593 .6261 .5207
.6639 1.0 .7571 .7162 .7214 .6387 .5457
.6942 .6539 1.0 .7992 .7816 .7323 .5971
.7554 .5971 .7346 1.0 .7451 .7224 .6097
.7687 .6679 .6929 .7604 1.0 .7365 .6107
.6365 .5306 .6636 .6381 .6579 1.0 .7454
.6302 .5729 .6381 .6458 .6812 .8376 1.0

Standard deviations
Group 1: 11.202 12.992 11.538 12.603 11.918 16.982 15.475
Group 2: 11.663 11.587 11.836 12.754 13.534 17.189 15.879

Means
Group 1: 24.664 25.643 25.087 30.241 28.535 54.174 69.075
Group 2: 31.508 31.559 31.723 34.949 34.026 60.531 72.280