An analysis of the 2003 HEFCE national student survey pilot data.

by

Harvey Goldstein

Institute of Education, University of London

h.goldstein@ioe.ac.uk

Abstract

The summary report produced from the first analyses of the HEFCE student 'course evaluation' survey contains suggestions for ways in which institutional comparisons can be presented. The present paper, from a member of the pilot project steering group, is intended as a further contribution to consultation and debate. It questions the practicality of the summary report suggestions and proposes more efficient data analysis procedures and more useable methods of presentation. It also urges further cautions in the interpretation of any results.

The subjects analysed are those that tend to be taken by large numbers of students. Even in these, there is not a great deal of significant separation that can be achieved between institutions; for many other subjects with smaller numbers of students in each institution we might expect that the data would yield few, if any, useful institutional comparisons. In the light of this and the various practical implementation difficulties, the value of the exercise, set against its cost and possible adverse side effects, is questionable.

Introduction

This analysis and commentary refers to the preliminary summary report of the pilot project that obtained data from final year students in 20 HE institutions who were asked to respond to questions about their courses. The summary report and a full description of the pilot can be found at http://iet.open.ac.uk/nss. The analyses produced 7 summary scales and it is these that are reported on. Table 1 provides the verbal descriptions that have been attached to the scales.

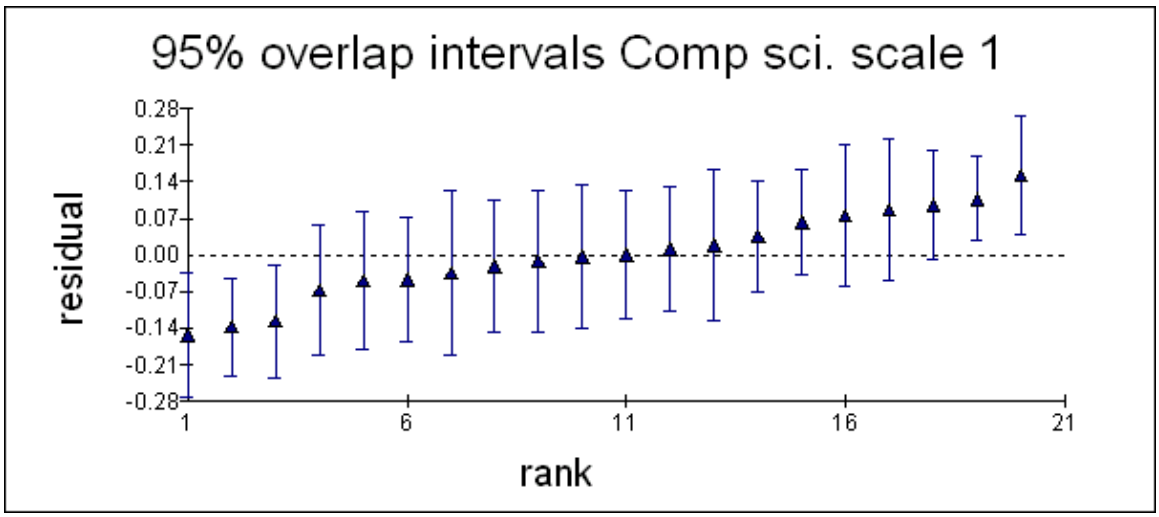
Table 1. Scales formed from student questionnaires.

Scale number	Description
1	Quality of staff teaching
2	Effectiveness of course feedback
3	Assesment criteria and arrangements
4	Effcet of course on 'generic' skills

5	Course workload balance
6	Study support and advice
7	Library, IT and other resources availability

The summary report presents tables and figures to illustrate how data from a full survey could be used and interpreted. The following two methods of presentation are used, although these may be modified in later reports.

The first presentation is in terms of 'overlap intervals' (Goldstein and Healy,1995). An example is the following for computer science course students



Essentially, if a single comparison is to be made between two institutional 'effects' (on the 'residual' axis then these are judged as significantly different at the 5% level if and only if the intervals overlap. This, for example there is no significant difference between the institution randked fourth from the bottom and any other institution. Note that this is only appropriate for a single comparison; for more than one pairwise comparison a multiple comparisons procedure is required.

The second type of presentation in the summary pilot report uses multiple comparisons for a hypothetical set of institutions. Three institutions are compared, pairwise, in the first table and five in the second. As the number of pairwise comparisons increases so some formerly significant differences become non significant (at the 5% level). The idea is to indicate the kinds of use a potential student might make of the data, when choosing among three of five institutions in terms of these rating scales.

The present paper extends the summary report in several ways. First, it is argued that the two methods of presentation described above suffer from certain drawbacks and secondly some extensive analyses using a multilevel model are presented as a more efficient alternative analysis method and leading to an alternative interpretation.

Problems with overlap intervals and multiple comparisons

The difficulty with presentation in terms of overlap intervals has already been alluded to. This is that it provides for a user to make just one pairwise comparison. While this may be what is required in some cases it will not be so in others. The multiple comparisons procedure, where there is more than one pairwise comparison, is intended to control the 'type I statistical error' at the 5% level and recognises this, but raises the further difficulty that a separate table would have to be provided for every possible number of comparisons desired. In practice there will be some individuals who may wish to carry out a large number of comparisons, and for these there will be few, if any significant differences Thus a table would have to be presented for every possible number of pairwise comparisons for every institution and this would be unmanageable.

An alternative is to provide an approximate scaling factor that could be applied to the pairwise overlap intervals. Thus, for 10 pairwise comparisons such a scale factor would be about 1.5. A further difficulty is that any user could access the tables more than once, carrying out new comparisons each time. If such a user made four comparisons then later made another six, the appropriate table would be based on ten comparisons and it would be quite complicated to implement a scheme that could cope with this. The alternative of using software, possibly on a web site, where users could input the set of comparisons they require, would encounter similar problems. The conservative option of choosing a maximum number of comparisons that a user would make in practice, would give intervals that were so large (for 40 institutions just over twice the length of the pairwaise intervals) that relatively few comparisons would be judged significant.

The following set of analyses is presented, therefore, as a more realistic and useful presentation of results pending any resolution of the difficulties described above. It provides a simple method for detecting whether any institution differs significantly from the overall average. In this way it should also help to avoid misrepresentation of the data in the form of crude 'league tables'.

Presentation based upon confidence intervals

Data for the 7 scales have been analysed using a 2-level model (Goldstein, 2003). Three subject groups are used, computer science, political and social studies and business studies since these cover most institutions. I am very grateful to John Richardson for making available the data and for helpful discussions with fellow members of the pilot project steering group.

Since some of the scale scores had skew distributions they are most appropriately analysed using a model that is based upon the scale scores treated as ordered categories, but this has not been done because of lack of time, although the software for carrying out such a (2-level) analysis is available and it is recommended that this analysis is done for a final version of any report. An approximation is to assign equivalent Normal scores to the scale scores. This has been done for a selection of scales but the results are little different from using the raw scale scores (see below for an example) and we have used the latter.

In the following analyses we shall present only some, typical, results: the complete set are in the appendix. Analyses were done in MlwiN (http://mlwin.com) and results are presented in the form of 'screen shots' from that software.

The 2-level model

For completeness the formal model is as follows:

$$y_{ij} = \beta_0 + u_j + e_{ij}$$
$$e_{ij} \sim N(0, \sigma_e^2), \quad u_j \sim N(\sigma_u^2) \qquad (1)$$

where *i* indexes students and *j* indexes institutions. If we wish to include covariates (x_h) then the model is extended as follows:

$$y_{ij} = \beta_0 + \sum_h \beta_h x_{hij} + u_j + e_{ij}$$
$$e_{ij} \sim N(0, \sigma_e^2), \quad u_j \sim N(\sigma_u^2) \qquad `(2)$$

Interest centres on the institution 'effects' u_j (referred to also as 'residuals') and the standard procedure is to estimate these using a linear predictor (see Goldstein, 2003 for a discussion). They are often known as 'shrunken' estimates since for small samples within an institution they are closer to the overall mean than estimates calculated simply by averaging the scores for each institution separately, as is done in the summary report. They posses certain technical advantages (such as minimising expected mean square) which suggests that they are to be preferred.

We can also obtain estimates of the standard errors of these effects and, assuming Normality (see below) this allows us to place confidence intervals around them. Such intervals need to take account of the fact that the between-institution variance is only estimated on the basis of 20 or so institutions. In practice this means that the usual 'plug-in' standard error estimates are too small. We have studied this for a few analyses (see below) and using the correct standard errors does slightly widen the intervals and these should be used in any final report. For now we do not use these.

Finally, it is often more informative to present results in terms of ranks (e.g. 1 to 20) with confidence intervals expressed in these terms also. We give an example or two below, but the general patterns are not much altered: in any final report it may be preferable to present in terms of ranks rather than using the original scale scores.

A full technical discussion of methods for estimating and presenting such rankings is given by Goldstein and Spiegelhalter (1996).

Results

Computer science

A complete set of analyses for each scale is provided in the appendix. Here we quote just scale 1 and scale 7 for illustration. We first give a screen shot of the fitted model and then one or more 'caterpillar' plots of confidence intervals.

Scale 1:

$$\text{scale1}_{ij} \sim \text{N}(XB,\ \Omega)$$

$$\text{scale1}_{ij} = \beta_{0ij}\text{cons}$$

$$\beta_{0ij} = 3.592(0.037) + u_{0j} + e_{0ij}$$

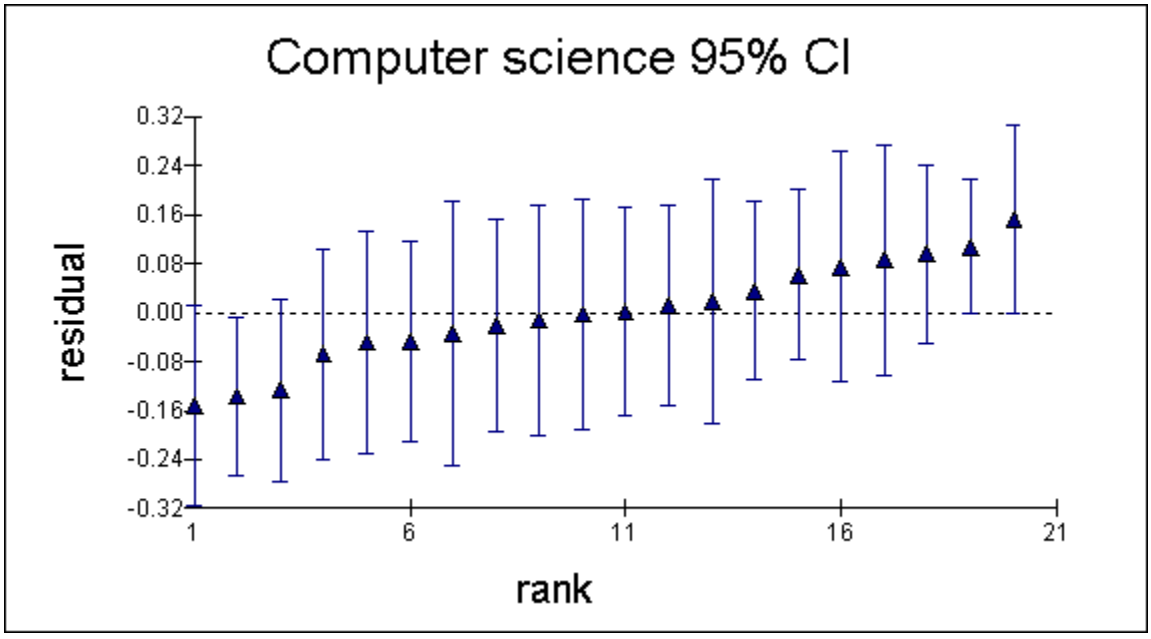
$$\begin{bmatrix} u_{0j} \end{bmatrix} \sim \text{N}(0,\ \Omega_u)\ :\ \Omega_u = \begin{bmatrix} 0.013(0.008) \end{bmatrix}$$

$$\begin{bmatrix} e_{0ij} \end{bmatrix} \sim \text{N}(0,\ \Omega_e)\ :\ \Omega_e = \begin{bmatrix} 0.526(0.023) \end{bmatrix}$$

$$-2*\text{loglikelihood(IGLS Deviance)} = 2311.685(1046 \text{ of } 1046 \text{ cases in use})$$

The percentage variance between institutions he's 0.013 divided by 0.013 + 0.526 which is 2.4%. The above plot shows the estimated residuals indicated by triangles, and ranked in order. The vertical bars give 95 per cent confidence intervals, and it is clear that only the two lowest ranked institutions can be separated statistically from the mean. The following caterpillar plot is based on the rankings of the institution residuals. Note that the mean is the model based population mean and is not estimated from the data.

Fig. 1



For comparison Fig 1a shows the 95% pairwise overlap intervals (Goldstein and Healy,1995) which provide paiwise significance tests for any given pair judged by whether intervals overlap (significant if not); these use a scaling factor of 1.45 to keep the overall significance level at 5%. The overall picture does not change substantially; most institutions cannot be separated on a pairwise basis. The top three can be separated from the bottom three, but not much else. There are, in all, 190 possible pairwise comparisons so this mean that about 95% of single pairwise comparisons ar enot judged as significant.

Fig 1a

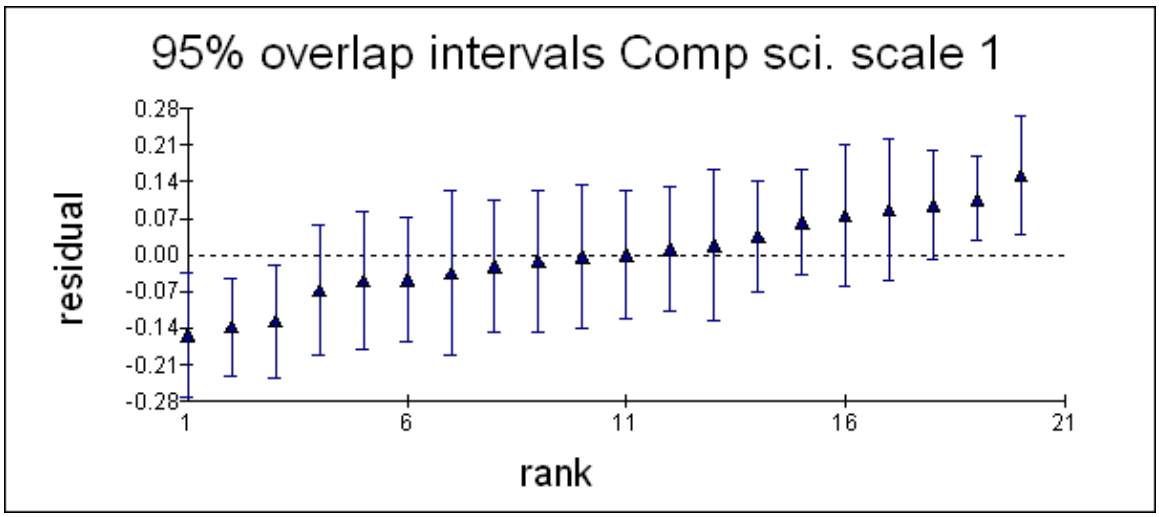
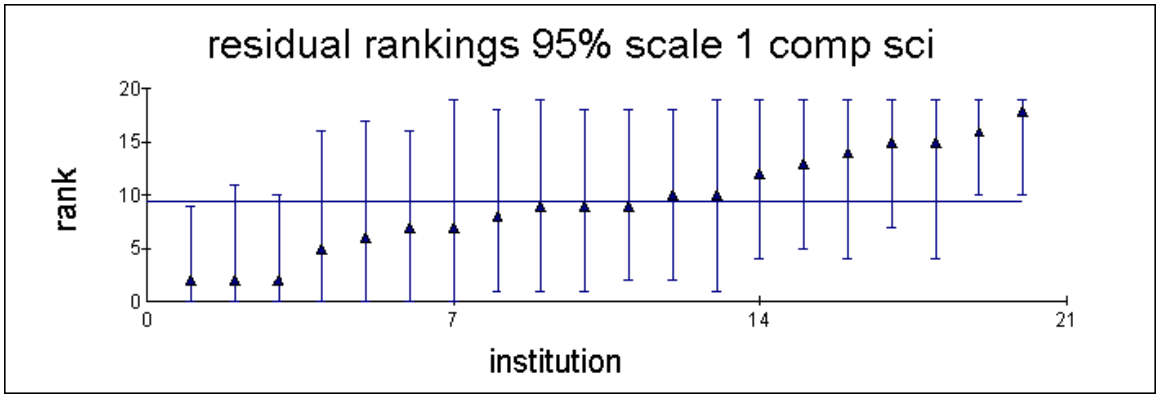
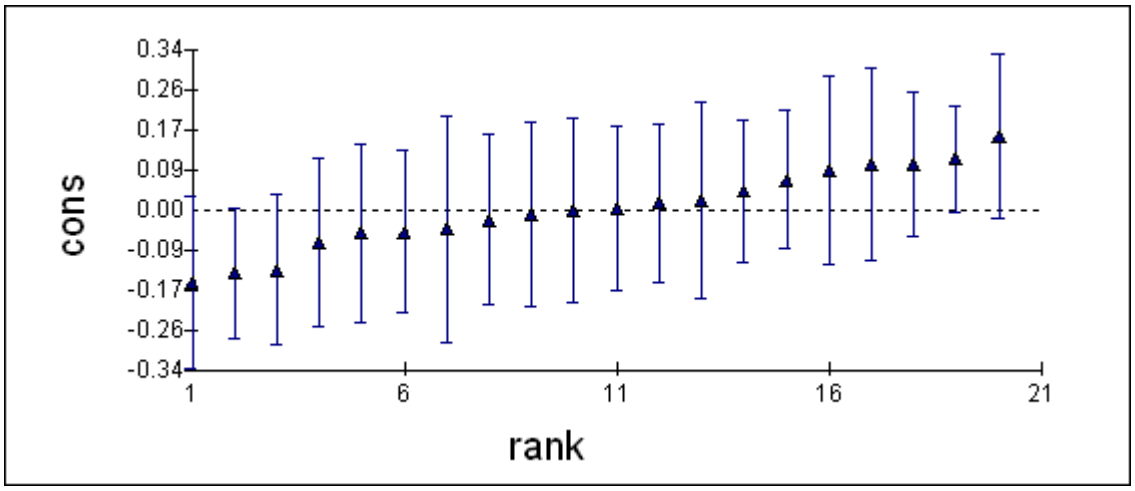


Fig. 2



Again we see that most of the institutions cover almost the whole range in terms of rankings. It is only just possible to separate four of the institutions as being above or below the mean rank. The next plot shows the residuals with 95 per cent intervals but allowing for the fact that the between institution variance is estimated from the model. We now see that the intervals are slightly longer with all of them overlapping the mean.

Fig.3



Scale 7:

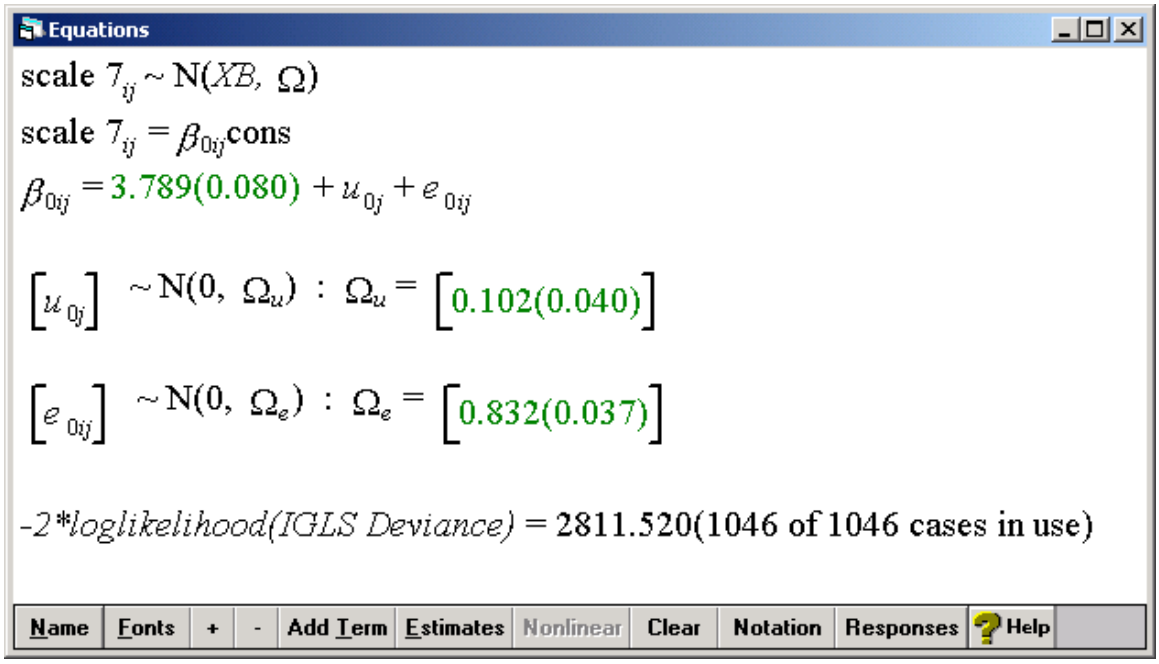
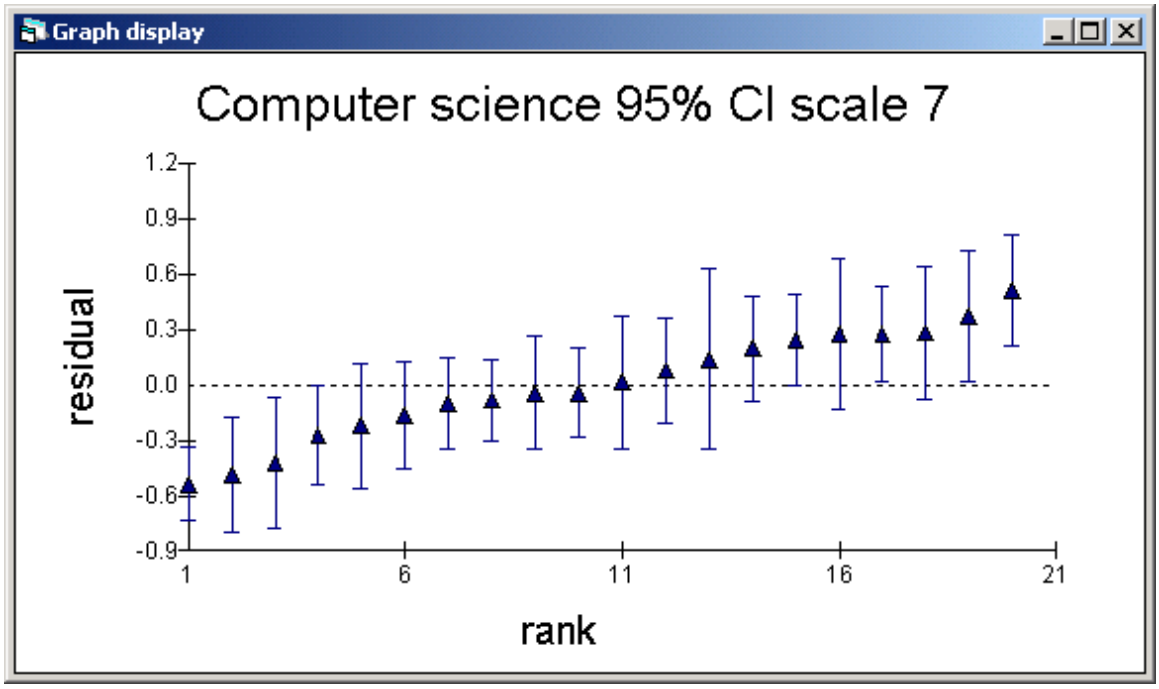
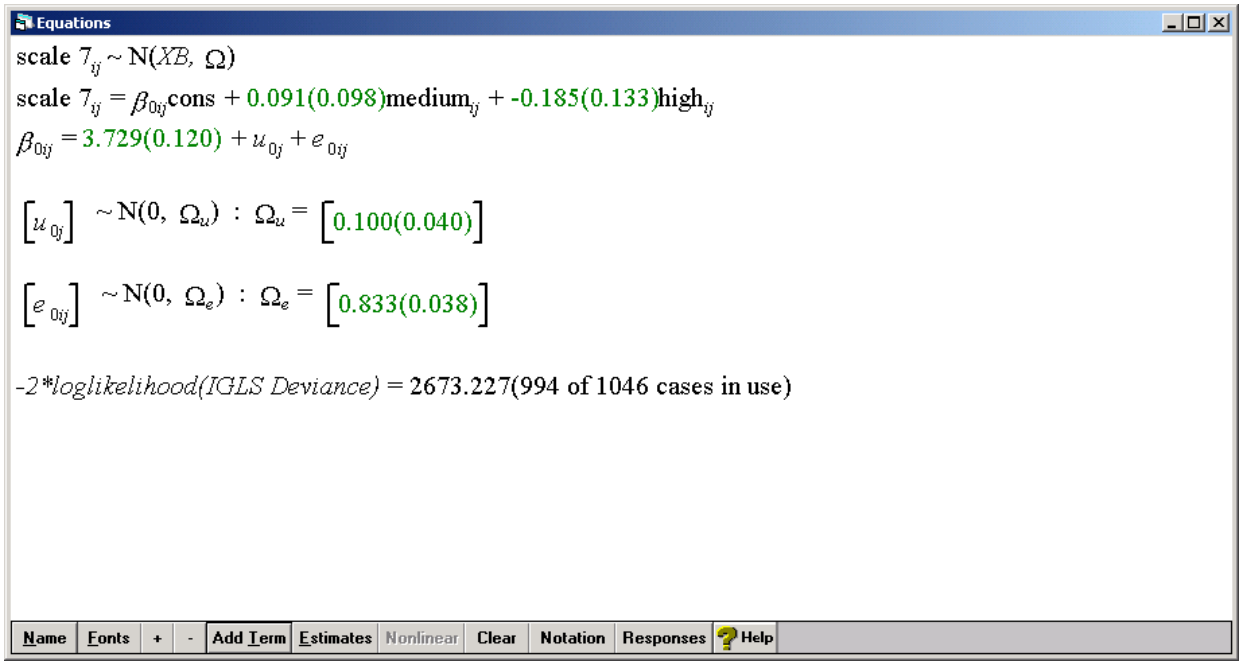

Equations
scale $7_{ij} \sim N(XB, \Omega)$
scale $7_{ij} = \beta_{0ij}$cons
$\beta_{0ij} = 3.789(0.080) + u_{0j} + e_{0ij}$
$[u_{0j}] \sim N(0, \Omega_u) : \Omega_u = [0.102(0.040)]$
$[e_{0ij}] \sim N(0, \Omega_e) : \Omega_e = [0.832(0.037)]$
-2*loglikelihood(IGLS Deviance) = 2811.520(1046 of 1046 cases in use)
Name
Fonts
+
-
Add Term
Estimates
Nonlinear
Clear
Notation
Responses
Help

Fig. 4



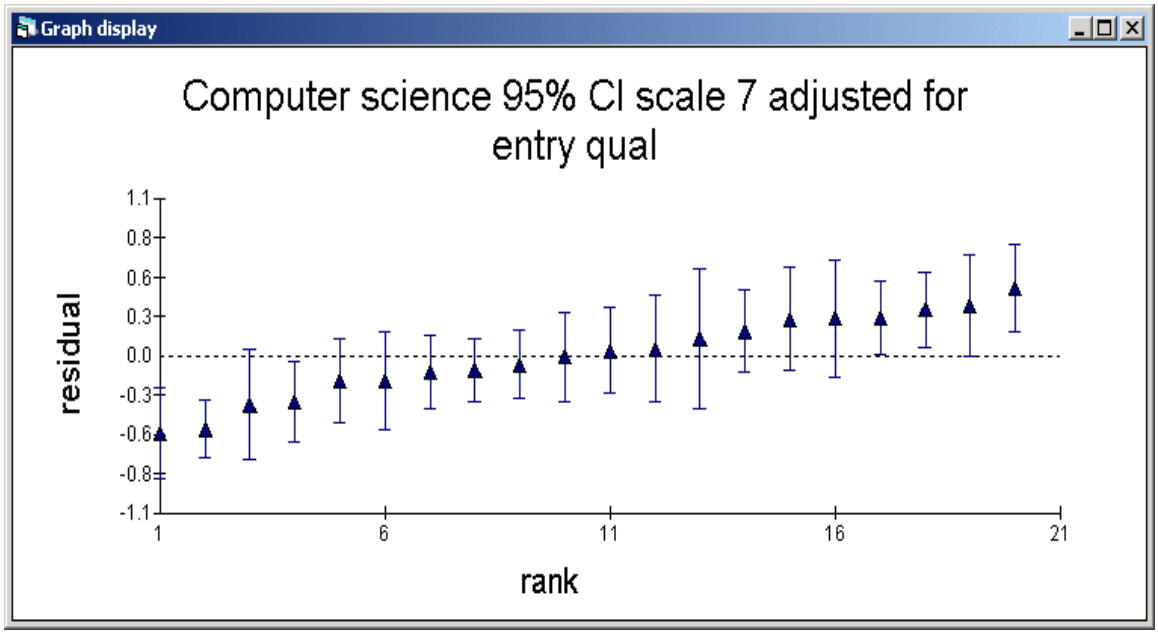
For scale seven 11 per cent of the variation is between institutions and there is slightly more separation. In this case the highest entry-level qualification is significant as shown in the following model.

Fig. 5



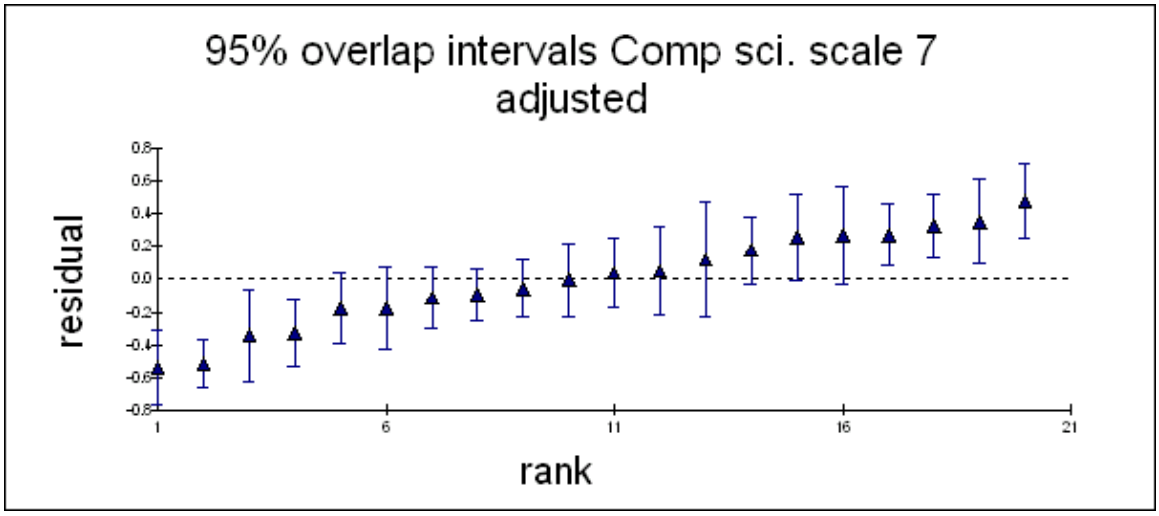
If we calculate institution effects for this model we obtain the following picture that is very similar to the above. A

Fig 6



For comparison, again, we present the 95% overlap intervals graph:

Fig 6a.



A rough calculation gives some 80% of institutions that cannot beseparated.

Political – social

Scale 1:

$$\text{scale1}_{ij} \sim \text{N}(XB,\ \Omega)$$

$$\text{scale1}_{ij} = \beta_{0ij}\text{cons}$$

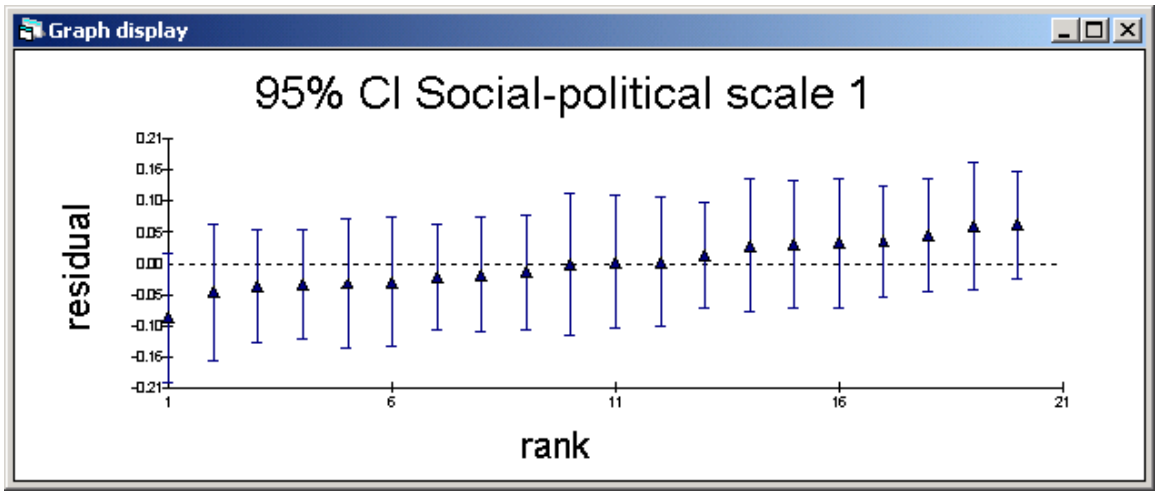
$$\beta_{0ij} = 3.903(0.023) + u_{0j} + e_{0ij}$$

$$\begin{bmatrix} u_{0j} \end{bmatrix} \sim \text{N}(0,\ \Omega_u) \ :\ \Omega_u = \begin{bmatrix} 0.004(0.003) \end{bmatrix}$$

$$\begin{bmatrix} e_{0ij} \end{bmatrix} \sim \text{N}(0,\ \Omega_e) \ :\ \Omega_e = \begin{bmatrix} 0.457(0.017) \end{bmatrix}$$

*-2*loglikelihood(IGLS Deviance)* = 3113.472(1511 of 1511 cases in use)

Fig. 7



We see for if this scale that there is no separation between institutions and they all are inseparable from the mean.

Scale 7:

Equations

$$\text{scale } 7_{ij} \sim \mathrm{N}(XB,\ \Omega)$$

$$\text{scale } 7_{ij} = \beta_{0ij}\text{cons}$$

$$\beta_{0ij} = 3.623(0.053) + u_{0j} + e_{0ij}$$

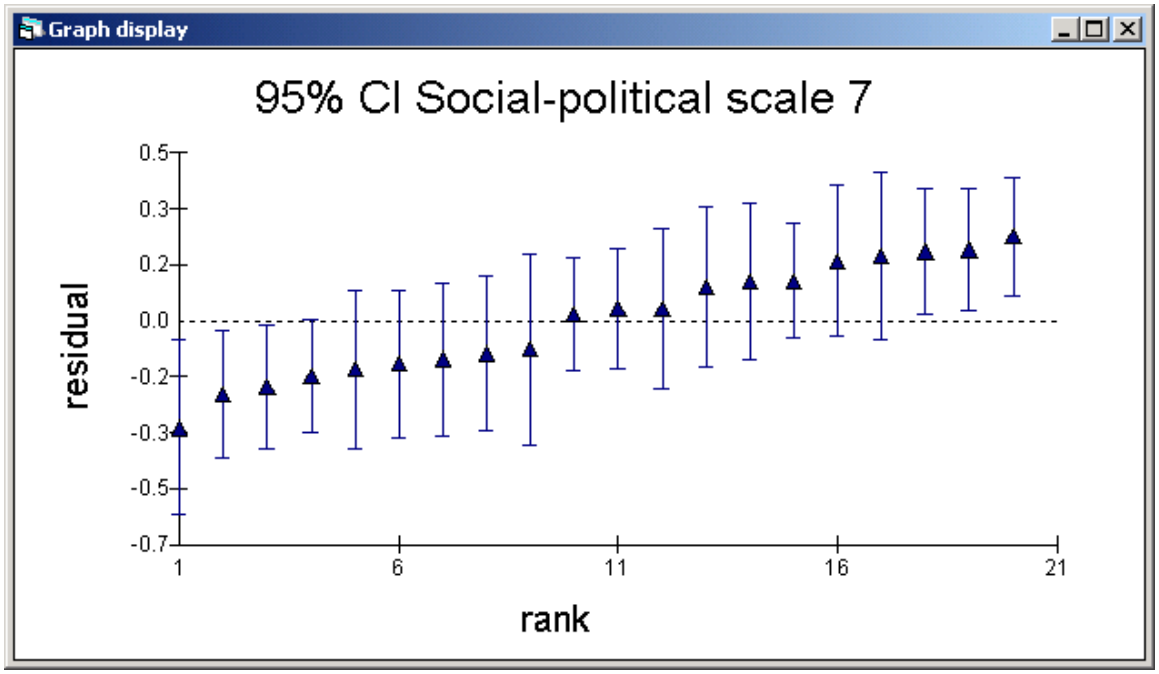
$$\begin{bmatrix} u_{0j} \end{bmatrix} \sim \mathrm{N}(0,\ \Omega_u) \ :\ \Omega_u = \begin{bmatrix} 0.041(0.018) \end{bmatrix}$$

$$\begin{bmatrix} e_{0ij} \end{bmatrix} \sim \mathrm{N}(0,\ \Omega_e) \ :\ \Omega_e = \begin{bmatrix} 0.933(0.034) \end{bmatrix}$$

$$-2*\mathit{loglikelihood(IGLS\ Deviance)} = 4210.581(1511 \text{ of } 1511 \text{ cases in use})$$

Name | Fonts | + | - | Add Term | Estimates | Nonlinear | Clear | Notation | Responses | Help

Fig. 8



For scale 7 for there is slightly more separation but again only a few institutions can be separated from the mean.

Business studies

Scale 1:

$$\text{scale1}_{ij} \sim \text{N}(XB,\ \Omega)$$

$$\text{scale1}_{ij} = \beta_{0ij}\text{cons}$$

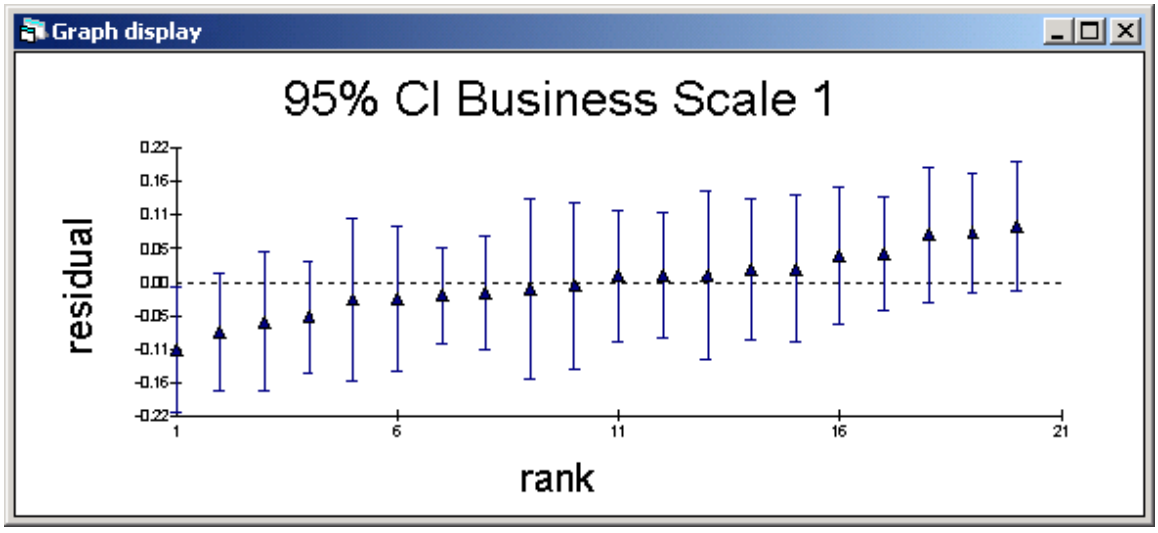
$$\beta_{0ij} = 3.708(0.024) + u_{0j} + e_{0ij}$$

$$\left[u_{0j}\right] \sim \text{N}(0,\ \Omega_u) \ :\ \Omega_u = \left[0.006(0.003)\right]$$

$$\left[e_{0ij}\right] \sim \text{N}(0,\ \Omega_e) \ :\ \Omega_e = \left[0.463(0.015)\right]$$

$$-2*\textit{loglikelihood(IGLS Deviance)} = 4083.170(1968 \text{ of } 1968 \text{ cases in use})$$

Fig 9



Full-scale one there is almost no separation from the mean.

Scale 7:

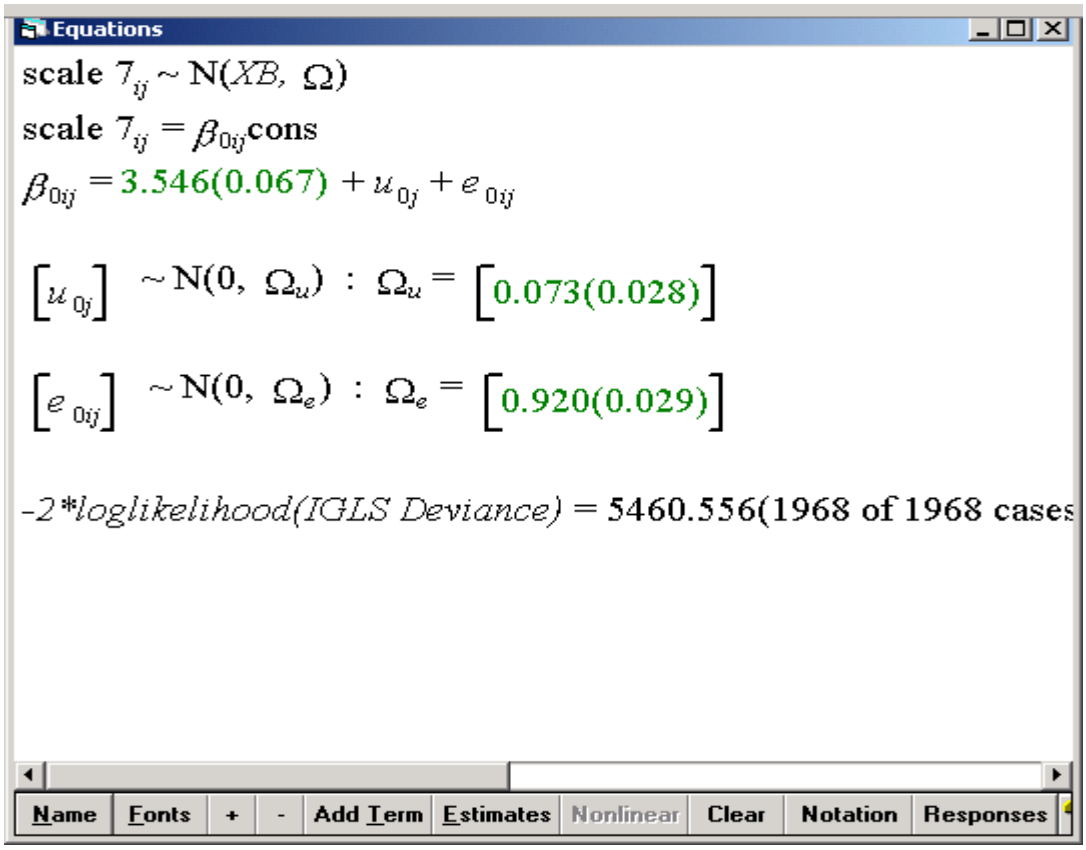
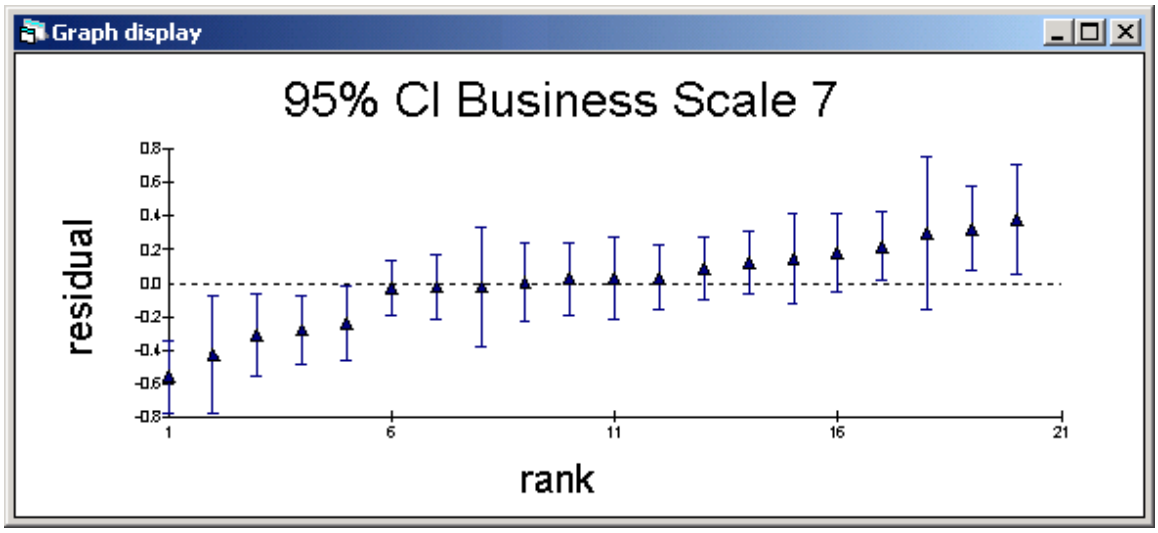


Fig 9



Again with scale 7 There is somewhat more separation among institutions.

Further thoughts

It might be supposed that combining the scales in various ways would produce more separation between institutions. This appears not to be the case in general, however.

Indeed, using the average over all scales we obtain the following results for computer science:

$$\text{total score}_{ij} \sim \text{N}(XB,\ \Omega)$$

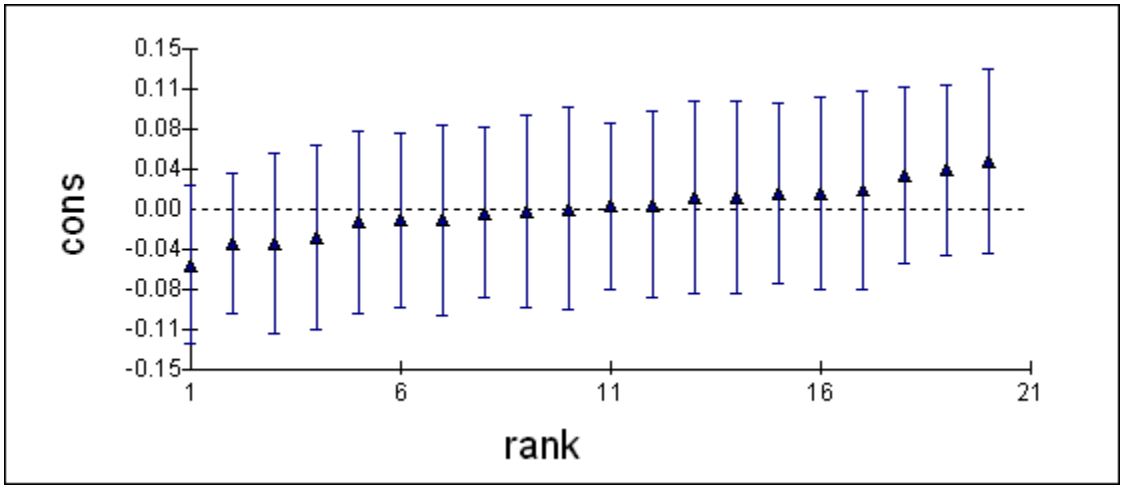
$$\text{total score}_{ij} = \beta_{0ij}\text{cons}$$

$$\beta_{0ij} = 3.401(0.022) + u_{0j} + e_{0ij}$$

$$\begin{bmatrix} u_{0j} \end{bmatrix} \sim \text{N}(0,\ \Omega_u) \ :\ \Omega_u = \begin{bmatrix} 0.002(0.003) \end{bmatrix}$$

$$\begin{bmatrix} e_{0ij} \end{bmatrix} \sim \text{N}(0,\ \Omega_e) \ :\ \Omega_e = \begin{bmatrix} 0.328(0.014) \end{bmatrix}$$

*-2*loglikelihood(IGLS Deviance)* = 1807.280(1046 of 1046 cases in use)



where there is no separation at all. In fact the correlation pattern among students seems to be different from the correlation pattern among the institution means (e.g. the correlation between scales 2 and 3 for students within institutions is 0.49 but between institutions is –0.14). This suggests that in carrying out factor analyses to determine scales a multilevel factor analysis should be done to allow the study of separate factors at the student and institution level.

When carrying out comparisons among institutions and judging significance or interpreting confidence intervals, we will obtain a certain number of 'positive' results purely by chance and this should introduce further caution into interpretations. Tentatively, based upon these preliminary analyses we may conclude that while in some cases some institutions appear to have ratings significantly better or worse than average, for most institutions for the subject areas we have looked at we cannot statistically provide meaningful rankings. In the special case of pairwise comparisons among institutions, even in the 'best' case some three quarters of the institutions cannot statistically be separated on the basis of a single pairwise comparison. Put another way, if

a pair of institutions is selecetd at random there is an 80% probability that they will not be significantly different.

A further issue is that we might well expect year-to year volatility for institutions and there is a strong case for using at least two years worth of data so that any such variation between years can be adjusted for.

While it is not the purpose of this paper to explore the consequences of introducing a performance monitoring system of this kind, nevertheless it is worth pointing out that such a system is unlikely to be free from adverse side effects. These might include effects such as a distortion of the teaching process in order to produce 'favourable' student evaluations, or an overdue concentration of, say, IT resources directed to the students due to fill in the questionnaire evaluations. Thus, if a system is introduced it will be important to monitor its progress and the possible emergence of such side effects.

While not all the possible institutional comparisons have been presented above, these are typical ones for the three subjects chosen. The full set can be obtained from the author: h.goldstein@ioe.ac.uk .

Finally, the subjects analysed so far are those that tend to be taken by large numbers of students. Even in these, there is not a great deal of significant separation that can be achieved between institutions; for many other subjects with smaller numbers of students in each institution we might expect that the data would yield few, if any, useful institutional comparisons. In the light of this and the various practical implementation difficulties, the value of the exercise, set against its cost and possible adverse side effects, is questionable.

References

Goldstein, H. (2003). *Multilevel Statistical Models. Third edition.* London, Edward Arnold:

Goldstein, H. and Healy, M. J. R. (1995). The graphical presentation of a collection of means. *J. Royal Statistical Society, A*. **158**: 175-7.

Goldstein, H. and Spiegelhalter, D. J. (1996). League tables and their limitations: statistical issues in comparisons of institutional performance. *Journal of the Royal Statistical Society, A*. 159: 385-443.