Multilevel modelling of PMETB data on trainee satisfaction and supervision

Chris McManus
March 2007

1. This final report on the PMETB trainee survey of 2006 is based on a finalised version of the SPSS data file[1]. A number of interim reports were submitted on earlier versions of the data, before final data cleaning and checking had been carried out, which clarified a number of aspects of the final analysis.

2. Although there are a number of possible dependent variables which can be considered, this analysis restricts itself to two composite measures, one a measure of overall satisfaction based on section H, items 2-6 of (henceforth '*satisfaction*'), and the other a measure of overall supervision based on items section C, items 1-5 (henceforth '*supervision*'). In each case the composites were provided by PMETB, and had been scaled so that they were in the range 0-100, with high scores meaning more satisfaction and better supervision. The choice of these variables was in part determined by the findings of the interim reports.

3. The main interest of the analysis was in a multilevel model (MLM) of differences between *deaneries*, between *training providers* (typically, but not always, hospitals or trusts), *Specialty Groups within training providers*.

4. In addition it was agreed to consider a number of background variables which were known to influence *satisfaction* and *supervision*:

   a. Specialty Group.

   b. Training grade

   c. Time in post

   d. Sex

   e. Route of responding to the questionnaire. In particular there were differences between those responding by paper questionnaires or electronically, and hence the six methods of responding were included separately.

   f. Year of qualification.

   Descriptive statistics for the dependent and background variables are provided in Appendix 1.

5. Of the 24,404 respondents included in the database, 23,267 had complete information on all of the data, and the present analysis is restricted to these individuals. The majority of those missing data were however missing only a single item. The most frequent item to be missing was the respondent's sex, which was not present for 512 trainees. The proportion of missing data is small, and hence the analysis was simpler without the use of formal processes of imputation, although that may be desirable at some future time.

[1] AllMergedNationalTraineeSurveyDataRedRemove_Feb07CMM.sav.

6. This analysis will principally be at the level of training provider, and it is worth noting that of the 755 separate providers, for 110 (14.6%) there was only a single respondent, 180 (23.8%) had only 2-5 respondents, 111 (14.7%) had 6-10 respondents, 140 (18.5%) had 11-30 respondents, 154 (20.4%) had 31-100 respondents, and 60 (7.9%) had more than 101+ respondents. The distribution is therefore very skewed, the mean of 30.8 respondents per provider being very different from the median of 9 respondents per provider, and the mode of 1 respondent per provider.

7. Multilevel modelling used *MLwiN 2.02*. The basic model considered here has four nested (hierarchical) levels of variance, *individuals*, who are within *Specialty Groups within providers*, who in turn are within *providers*, who are within *deaneries*.

8. *Differences between Speciality Groups*. It should be noted that there are large differences between Specialty Groups in mean satisfaction and mean supervision, and therefore Specialty Group has been entered as a fixed effect in all of the analyses. Appendix 2 provides a detailed breakdown.

9. **Satisfaction**. The first part of this report will analyse the satisfaction scores of the trainees.

10. *Complex variance modelling*. *MLwiN* can not only model variance at different hierarchical levels but can also model error variances in terms of other variables including the dependent variable itself. Preliminary analyses of the data organised by Specialty Group found that not only did mean satisfaction vary between Specialty Groups, but so also did the variances (standard deviations), those Specialty Groups with higher mean scores having lower variances (see Appendix 2). That is almost certainly due to a ceiling effect, the most extreme theoretical case being where everyone within a Specialty Group is entirely happy, so that the mean score is at the maximum of 100 and, of necessity, there is a standard deviation of zero. It therefore seemed likely that error variance at t level 1(*individuals*) would be correlated with satisfaction. This is what *MLwiN* causes "complex variance".

11. The possibility of complex variance was explored firstly using a simple model in which variance was allowed to vary at each of the four levels (see figure 1a). The model was then extended so that at level 1, the variance was a linear function of the composite satisfaction variable (see figure 1b)[2]. The goodness of fit was substantially improved, from 190,918 to 187,985, and the linear effect of satisfaction on variance was highly significant (-2.692, SE=.062), the negative coefficient indicating, as expected, that variance decreases as satisfaction increases[3]. Complex variance was included in future models based on this linear trend of variance upon satisfaction at the individual level. The overall effect, as seen in figure 1, is to reduce somewhat the variance at higher levels of the model, but also to reduce the standard errors of those variances, so that significance levels are little affected.

12. The model was then refitted to include fixed effects of a range of background variables. In particular, Specialty Group was fitted, since it was clear that trainees in some Specialty Groups are more satisfied than others (and the mix of Specialty Groups also differs between providers and between deaneries). The important feature,

[2] Note that for technical reasons, this was done by creating a second variable called *satisfac2*, which was identical to *satisfac*.

[3] The model was explored a) in which at level 1 there was a quadratic effect upon error variance, and b) linear effects of satisfac were allowed on levels 2, 3 and 4 variance; in each case estimates failed to converge.

which must be emphasised, is that as a result differences between providers or, in particular, differences between Specialty Groups within providers, cannot be due to differences in average satisfaction of Specialty Groups, or to a different Specialty Group mix within providers, as mean differences between Specialty Groups have already been taken into account by including Specialty Group as a fixed factor.

13. Several other fixed effects were also included which might have an impact on overall satisfaction. These included, route of return of the questionnaire (there being a suggestion that the more anonymous the method then the higher likelihood of errors and other socially negative events being reported), sex, year of qualifying (linear trend), time in post, and grade. The overall model is shown in figure 2, and is significantly improved over that in figure 1b, showing that the fixed effects do have an important influence. Since they are of little interest for present purposes, they will not be explored further here, although there is probably much of interest in them.

14. The most interesting aspects of figure 2 concerns the variance at the higher levels, and it is clear that there is significant variance at level 2 (*Specialty Groups within providers*; 12.591, SE = 1.046, z=12.04, p<<.001) and level 3 (*providers*; 2.140, SE .653, z=3.27, p=.0011), although the variance at level 4, *deanery*, is not significant (.779, SE .408, z=1.909, p=.056). Although the deanery differences do not quite reach conventional levels of significance, they have been left in the final model so that there is no risk of provider differences being confounded with them.

15. *A note on interpreting variances*. Although the variance at the level of the deanery is less than that at the level of the provider, and that at the level of providers is less than that at the level of Specialty Groups within providers, that should not be interpreted as meaning that Specialty Groups within providers have greater impact than do providers. Variance at aggregated levels is always less than at lower levels (and so for instance the total variance of income *within* countries such as the US or UK is always greater than that of the variance in average income *between* countries. The likes of Bill Gates always means that the variance within countries is higher than between countries, the US and UK mean incomes for instance differing by thousands of dollars on average, whereas income in the US has a much wider distribution, with at least 400 billionaires). When variance is present at higher levels in the model it arises because average levels within some groups are higher than in others, and hence it reflects effects which affect *many* people.

16. The intention of the trainee survey was to assess differences at the level of the provider and individual Specialty Groups within providers. The remainder of the analysis will therefore be concerned with these levels, and in particular the provider level. As already noted above, many providers have only a single respondent, and over a half (53.1%) have ten or fewer respondents. The question of whether differences are significant and meaningful therefore needs careful consideration. The analysis will concentrate on the provider level in the first instance.

17. *Differences between providers*. Figure 3a shows a plot of the effects for each of the providers (technically known as residuals), with providers ordered from smallest to largest. Figure 3b shows the same effects but with error bars (1.96 SEs), equivalent to 95% confidence intervals. The plotting of error bars in such graphs is complex, because they implicitly make multiple comparisons. In this and other graphs the bars should only be used to compare a single residual with the population mean (shown as a dashed line). Appendix 3 considers in more detail the difficult question of plotting error bars within MLM. Figure 3c plots residuals against normal deviates, and it can be seen that the line is approximately straight, implying that differences between providers are normally distributed to a first approximation. Although the differences

in figure 3a may seem large, all but one of the errors bars, at the top end, includes the population average.

18. In order to explore the provider level effects further, the residuals from *MLwiN* were merged with a version of the original SPSS file aggregated at the provider level.

19. Figure 4a shows a scattergram of the provider effect (residual) calculated by *MLwiN* (vertical), in relation to a simple, raw average of the satisfaction scores of the individuals in the provider (horizontal). In addition, data are plotted separately according to the number of respondents for each provider. A number of things are noticeable about the graph:

    a. The raw means on satisfaction have less variability when there are more people responding, as will be expected. The most variance is shown with the red points which correspond to providers with a single individual. That is a standard and expected finding.

    b. In contrast, the fitted effects from *MLwiN* show more variability when there are more respondents, as they are proper estimates at the provider level. Multilevel modelling takes into account all of the information from providers, even when there is a single respondent, and uses it all to fit a model. In effect, instead of assuming that when there are fewer respondents in a group there is less information about variances, MLM assumes that the variances within any one group is similar to that within other groups, and hence variances from all groups contribute to a knowledge of variance within a particular group. This assumption is similar to that in the more familiar one-way analysis of variance.

    c. As a result of the influence of number of respondents, the regression lines are much steeper where there are larger numbers of respondents than smaller.

    d. Finally, it should be noted that overall there is a reasonable correlation (r=.428) between the raw estimates and the *MLwiN* residuals, which is both reassuring in that it is not zero (and would be surprising if it were so), and also emphasises the additional benefit gained from using *MLwiN* (for otherwise the correlation would be one).

20. *MLwiN* provides a comparative standard error for each residual, and from this can be calculated a z-score for the difference of the residual from the population average (which by definition is zero), and then can be converted to a significance (probability) level. Of the 755 z-scores, only 1 (0.1%) was significantly different from the population average with $p<.05$. Figure 4b shows a plot of significance (vertical) in relation to the number of respondents (horizontal). The mean (log) significance level shows a negative correlation with number of respondents (r=-.422), larger studies showing somewhat more significant differences (which is hardly unexpected). There are two horizontal lines on figure 4b, a solid one at p=.05 and a dashed one at a Bonferroni corrected significance level of 0.05/755 = .000066. This correction may be overly conservative, but it is also likely that $p<.05$ is overly liberal.

21. Figure 5a shows estimates of the *MLwiN* effect size (horizontal), plotted against significance level calculated from the comparative standard error (vertical), with different colours showing the number of respondents in six groups. In addition, there are two horizontal lines, a solid one at the 5% level of significance, and a dashed line at the Bonferroni corrected level. As mentioned earlier, only one of the points is below the 5% significance line, and it is far removed from the Bonferroni corrected significance level.

22. Figure 5b, for interest and to show the advantages of multilevel modelling, shows the raw, unadjusted mean satisfaction score for each provider, in relation to the level of significance in the multilevel model. As before there is only one blob below the 0.05 level, but interestingly it is almost precisely at the population mean for satisfaction. The implication is that adjustment for a range of background variables has allowed the relatively high satisfaction of those respondents to become clear.

23. *Differences between Specialty Groups within providers*. In just the same way as one can analyse residuals at the provider level, as described above, so one can also analyse residuals at the level of Specialty Groups within providers (level 2).

24. Overall there are 2,687 combinations of Specialty Groups within providers, but for 695 (25.9%) there was only a single respondent, for 800 (29.8%) there were only 2-5 respondents, for 503 (18.7%) there were 6-10 respondents, 556 (20.7%) with 11-30 respondents, 129 (4.8%) with 31-100 respondents, and only 4 (0.1%) with more than 100 respondents. Overall, therefore, 74.4% of Specialty Groups within providers had ten or fewer respondents, with a very skewed distribution, the mean being 8.66 respondents, while the median was 5 and the mode was 1.

25. Figure 6a shows the effects (residuals) for each Specialty Group within provider, ranked from smallest to largest. Although the error bars are wide, and the numbers of bars makes it difficult to see what is going on, a number do not include the population average. These will be considered in more detail in a moment when looking at figure 7a. Figure 6b plots the residuals against a normal deviate, and the line is almost entirely straight, suggesting that the population distribution is a good approximation to normal.

26. Figure 7a shows the residuals (horizotnal) plotted against their significance level (vertical), and figure 7b shows a similar graph but based on the raw, unadjusted mean satisfaction values (horizontal). The horizontal lines show a raw, unadjusted 5% level of significance, and a Bonferroni adjusted level of 0.05/2687=.000019 (note that N has changed so the critical P has also changed). None of the effects reaches the Bonferroni adjusted level, although there are 53 (2.0%) cases which are beyond the 0.05 level, 30 with high values and 23 with low values. Whether the differences described here are truly significant is a difficult question, because of the problem of multiple testing. None has achieved the conventional Bonferroni criterion, which must throw doubt on the differences, although the Bonferroni correction is probably overly conservative.

27. *Benjamini and Liu's modification of the Bonferroni correction*. Although a number of Specialty Groups within providers showed raw, uncorrected significance levels below the 5% level, none achieved significance after a conventional Bonferroni correction for multiple testing. However the conventional Bonferroni is perhaps too conservative in such situations. Nevertheless, raw significance levels are undoubtedly far too liberal. As Benjamini et al (1) have said, what is needed is a method that acknowledges,

> "the importance of controlling for the treacherous effect of multiplicity, while not being overly conservative".

28. The problem is a generic one in many areas of technical research, as for instance in screening of drugs against batteries of behavioural outcomes, searching multiple genes for quantitative trait loci, looking at the outcome of gene arrays, or examining the millions of pixels produced by a typical fMRI scan. The purpose of many such studies is *discovery*, and while a Bonferroni correction controls the rate of false positives such that it is, in effect, zero (so that the FWER – the family-wise error rate

is 5%), this is not what is required when discovery is the aim of the study, and a certain rate of false positives can be tolerated. Benjamini and Hochberg (2) have therefore emphasised the importance of what is called the false discovery rate (FDR), and suggest that also needs controlling. In particular they argue that if one is willing to use, say, a 5% significance rate (alpha), one should also be willing to accept that 5% of those discoveries will actually be non-significant. The power of the procedure is therefore increased, at the cost of increasing the proportion of false positives.

29. A simple method of modifying the Bonferroni correction for the FDR was described by Benjamini and Hochberg (2), and modified by Benjamini and Liu (3;4), as described in a simple way by Benjamini et al (1). The method has the advantage of only requiring the significance levels of each of the N tests that has been carried out, and is therefore generally applicable, and can be straightforwardly carried out in a spreadsheet program. It should however be noted that more sophisticated methods are available within the context of multilevel modelling, as described by Afshartous and Wolf (5), although they are technically difficult. It is not however clear that they can be applied to more than two levels, and they do not run within conventional MLwiN or MLM software.

30. Here Benjamini and Liu's (BL) method is applied to the 2,687 residuals obtained at the level of Specialty Groups within providers. which gives a very straightforward result; none of the residuals was statistically significant at a corrected 5% level. The analysis was also carried out for the 689 Specialty Groups within providers with more than ten respondents, and also for the 133 Specialty Groups within providers with more than thirty respondents (so that the adjustment was based on a lower N than for the 3,216 residuals). However in neither case did any Specialty Group within provider achieve significance.

31. In order to check the workings of the BL method within the context of a multilevel model, I have re-analysed the (much analysed) LEA data described in the MLwinM manual, and analysed also by Afshartous and Wolf (5). Following the method described in the manual, 31 of the schools are “significant” at the .05 level. However this method is valid only for a single comparison. Afshartous and Wolf’s complex calculations give 17 significant results for StepM (see their tables 1 and 2). Applying the BL method to the same data finds 16 schools significantly different from the population average. The BL method appears to be broadly equivalent to more sophisticated methods, and therefore can probably be relied on as a good first approximation.

32. In conclusion, after accounting for multiple significance testing because of large numbers of different Specialty Groups within providers, and using a better criterion than the simple Bonferroni correction, there is no reason to believe that any of the Specialty Groups can convincingly be regarded as substantially different from others. The Afshartous and Wolf correction has not yet been applied to these data, as the software is not yet available.

33. *A simpler model with only provider level variation*. Since there is no evidence of variation at the level of deaneries, and there has to be a concern that power has been lost by having variation at the level of both providers and Specialty Groups within providers, not least since the modal number of respondents in each is one, a simpler multi-level model was fitted in which variation was only present at the level of the individual and the provider. Once again there was significant variation between providers (variance = 4.665, SE = .605), and 41 of the providers reached the conventional .05 significance level (22 above the mean and 19 below the mean). However the most significant effect only had a significance level of p=.0000727,

which did not reach the Bonferroni corrected level of .05/755=.000066, and neither was it significant with the Benjamini and Liu method. The conclusion has to be that fitting deaneries, providers and Specialty Groups within providers has not resulted in any serious loss of power for detecting differences between providers.

34. **Supervision**. The analysis of differences in perceived supervision level will parallel those of satisfaction, and can be reported far more briefly.

35. As before a model was fitted with four levels, which included complex variation at level 1, the latter being highly significant (see figure 8; component = -2.963, SE=.053). In addition all other background variables were included as before, with the exception of year of qualification, the inclusion of which, for reasons that are not clear, prevented convergence of the model.

36. As in the model for satisfaction, there was highly significant variation at the level of providers (.869, SE=.310, z=2.803, p=.0051), and Specialty Groups within providers (4.618, SE=.513, z=9.002, p<<.001), although there was not a significant effect of deanery (.173, SE=.127, z=1.377, p=.168).

37. In the raw data at the respondent level, there is a correlation of .482 between supervision scores and satisfaction scores. A model of supervision scores was therefore fitted in which satisfaction was entered as a covariate. The effects of provider (.742, SE .248) and Specialty Group within provider (3.039, SE=.397) were still highly significant (although again there was no effect of deanery; effect=.091, SE=.085), indicating that the effects of supervision are independent of those of satisfaction. A similar result was found when the data for satisfaction were re-analysed with supervision as a covariate.

38. Figure 9a shows the residuals at the level of the provider with their 95% confidence intervals for a single comparison to the mean. As with the data for satisfaction, there are only two confidence intervals that do not overlap the mean, and these only just reach significance (p=.029 and .031) before correction, and are not even close to a Bonferroni corrected significance level (p=.05/755=.000066).

39. Figure 9b shows residuals at the level of the Specialty Group within providers, again with 95% confidence intervals. Of the 2687 residuals, only 31 reach a raw, uncorrected significance level of p<.05 (10 low and 21 high), and none is significant using the Bonferroni corrected method or the method of Benjamini and Liu.

40. **Conclusions.** The conclusions from this multilevel model of satisfaction and supervision are relatively straightforward.

    a. The multi-level modelling shows very clearly that there are differences in the satisfaction and supervision of respondents, and that there is variance both at the level of *providers* and of *Specialty Groups within providers*, although there is no variance at the level of *deaneries*. The differences do not reflect differences in trainee mix in terms of Specialty Group, sex, grade, type in post, years qualified, or route of responding to the questionnaire. In interpreting this result it should be remembered that just as “half of all doctors are below average” (6), so half of all providers and half of all Specialty Groups within providers will be below average, although the data suggest that the variation around the average is greater than would be expected due to sampling variation alone.

    b. The overall differences between providers and within providers are statistically robust (i.e. considering the data as a whole), and certainly require explanation. In the absence of compositional variables describing providers or

the Specialty Groups within providers, it is not possible to get any further in explaining the differences, and in a future study it would be desirable to collect such measures – one thinks perhaps of measures such as senior-junior ratio, doctor-patient ratio, work-mix measures, or whatever.

c. Although the analysis provides little doubt that providers and Specialty Groups within providers do show significant differences, it is also clear that the data are not sufficiently robust to be able to identify a particular subset of providers or Specialty Groups within providers that are particularly underperforming or are performing significantly better than average, very few providers and Specialty Groups being outside the conventional 0.05 level, which itself is not of course corrected for the large numbers of providers.

d. The failure to detect specific poorly performing providers or Specialty Groups within providers reflects the well known problem that measures are often more than adequate for demonstrating significant differences between groups, but nevertheless can say remarkably little about measures within individuals. A classic example in psychology concerns intelligence tests, where scores on a couple of a hundred or so people can readily show mean differences between groups of only a few points, and yet an individual's intelligence cannot accurately be identified to within about five to ten points either side of their true score (and that is despite intelligence tests often having relatively high reliability of measurement, of about 0.95).

e. The failure to find significant differences at the level of individual providers, despite there being clear overall evidence of differences between providers, can be analysed in terms of two factors:

   i. *Reliability of measures*. If individual measures are unreliable then little can be found out about how they differ between groups. The five measures of the satisfaction scale are relatively reliable, with a Cronbach's alpha of .896, which is acceptable. However the five items that comprise the supervision scale are less reliable, having a Cronbach's alpha of only .530, which is adequate for finding group differences with a large sample size, but is not satisfactory for identifying individual providers that are problematic.

   ii. *Sample size*. Although the overall sample size for the study is large, with over 23,000 respondents with complete data, the problem arises that there is also a very large number of providers, 755, each of whom on average has three or four Specialty Groups, and some of which have many more, making 2687 Specialty Groups within providers. The result is that the modal number of respondents per provider and per Specialty Group within provider is precisely one. That is simply too few to be able to expect to find significant results. Even when the analysis is restricted to Specialty Groups within providers with 10 or even 30 respondents, there is still no evidence of significant differences between the groups. Of course that may also reflect the fact that larger organisations tend to have a more systematic approach to managing their trainees, and hence are less likely to be outliers.

41. *The future*. The primary problem for identifying poorly performing providers from these data, is that the sample size is too small. In addition even composite measures such as supervision are relatively unreliable. If the main aim of the study is to find poor providers, then the deep structural problem of having only a few respondents per

provider is not readily going to be overcome. Even so, it should always be remembered that the ultimate test of statistical significance is whether or not a result replicates in an independent sample. If a second study were to be carried out, using precisely the same measures in the same providers, then if a similar rank ordering is found on the second occasion, and in particular the especially good or the especially bad on the first occasion showed the same effect on the second occasion, then that might be sufficient evidence both of reliability and a basis for action. My prediction however is that such a result will not be found, the extreme results being dominated at present by the smallest samples, and hence being the least reliable.

42. Full replication is expensive, tedious, and also requires waiting so that another cohort of respondents comes through. If some form of validation of the present results is required, then a more experimental form of analysis may be better. If one were to take, say, three groups of providers (or perhaps Specialty Groups within providers), one at the bottom end of the distribution, one at the top, and the third around the mean, with, perhaps 10 in each group, then those 30 providers could be inspected by individuals blind to the group assignment. If experienced trainers can identify the good, average and poor providers, then clearly the measures are telling us something significant, whereas if they cannot, then probably the measures mostly consist of random measurement error.

Acknowledgments. I am extremely grateful to Daniel Smith for his assistance in the analysis of these complex data.

Appendix 1. Descriptive statistics and histograms for variables.

Speciality

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	2 Anaesthetics	2519	10.8	10.8	10.8
	3 Emergency Medicine	1191	5.1	5.1	15.9
	4 General Practice	1804	7.8	7.8	23.7
	5 Medicine	5705	24.5	24.5	48.2
	6 Obstetrics and Gynaecology	1702	7.3	7.3	55.5
	7 Occupational Medicine	35	.2	.2	55.7
	8 Ophthalmology	450	1.9	1.9	57.6
	9 Paediatrics and Child Health	2248	9.7	9.7	67.3
	10 Pathology	524	2.3	2.3	69.5
	11 Psychiatry	2336	10.0	10.0	79.6
	12 Public Health	87	.4	.4	79.9
	13 Radiology	817	3.5	3.5	83.5
	14 Surgery	3849	16.5	16.5	100.0
	Total	23267	100.0	100.0	

grade Which grade is your current post?

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	1 SHO	8994	38.7	38.7	38.7
	2 SpR	10844	46.6	46.6	85.3
	3 GPVTS – SHO in hospital	1577	6.8	6.8	92.0
	4 GPVTS SHO in GP practice	229	1.0	1.0	93.0
	5 GPR	1623	7.0	7.0	100.0
	Total	23267	100.0	100.0	

sex Please indicate your sex.

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	1 Male	13157	56.5	56.5	56.5
	2 Female	10110	43.5	43.5	100.0
	Total	23267	100.0	100.0	

timepost How long have you been in your current post? (Current post in rotation, not overall training period).

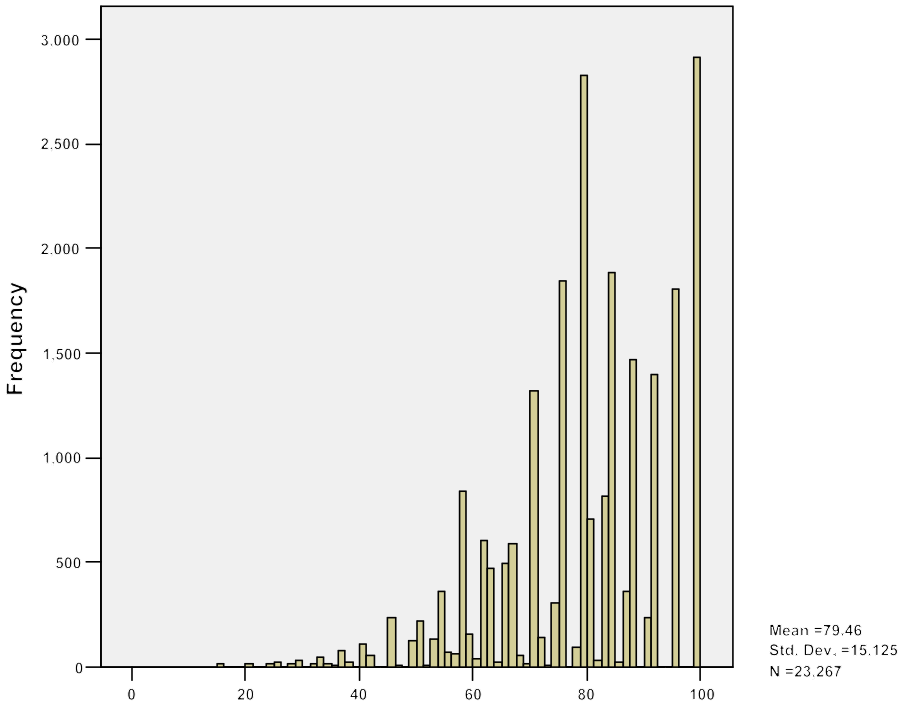
		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	1 Less than 1 month	457	2.0	2.0	2.0
	2 1 to 3 months	2082	8.9	8.9	10.9
	3 More than 3 months, but less than 6 months	9242	39.7	39.7	50.6
	4 6 months and over	11486	49.4	49.4	100.0
	Total	23267	100.0	100.0	

yrqual

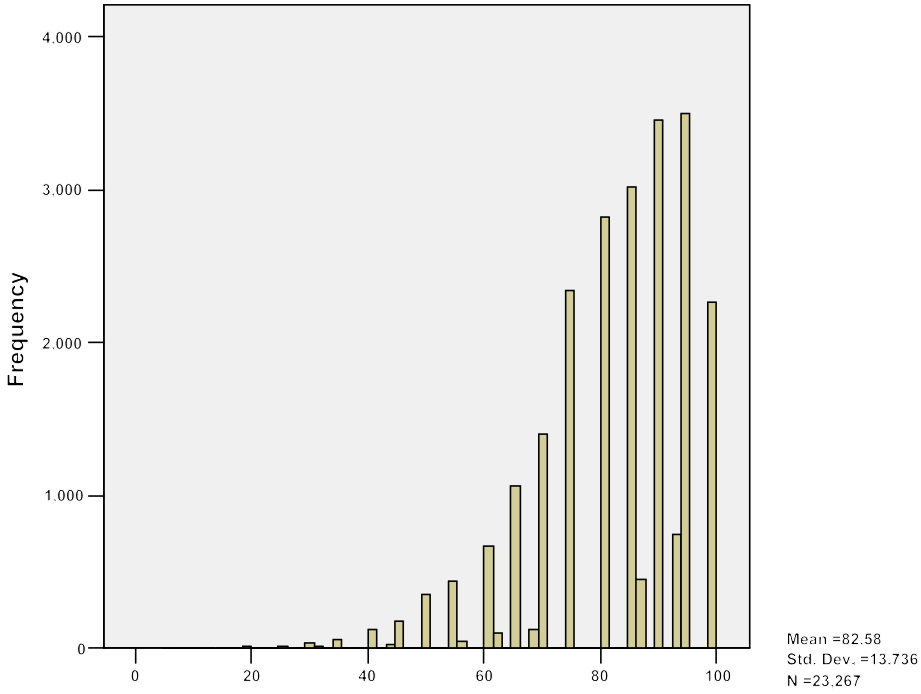
		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	1979	61	.3	.3	.3
	1980	23	.1	.1	.4
	1981	26	.1	.1	.5
	1982	42	.2	.2	.7
	1983	53	.2	.2	.9
	1984	76	.3	.3	1.2
	1985	82	.4	.4	1.6
	1986	106	.5	.5	2.0
	1987	134	.6	.6	2.6
	1988	161	.7	.7	3.3
	1989	216	.9	.9	4.2
	1990	308	1.3	1.3	5.5
	1991	397	1.7	1.7	7.2
	1992	542	2.3	2.3	9.6
	1993	637	2.7	2.7	12.3
	1994	1009	4.3	4.3	16.6
	1995	1279	5.5	5.5	22.1
	1996	1531	6.6	6.6	28.7
	1997	1768	7.6	7.6	36.3
	1998	2039	8.8	8.8	45.1
	1999	2076	8.9	8.9	54.0
	2000	2270	9.8	9.8	63.8
	2001	2199	9.5	9.5	73.2
	2002	2441	10.5	10.5	83.7
	2003	2288	9.8	9.8	93.5
	2004	1473	6.3	6.3	99.9
	2005	30	.1	.1	100.0
	Total	23267	100.0	100.0	

route

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	1 Web	6905	29.7	29.7	29.7
	2 Paper scanned	7757	33.3	33.3	63.0
	3 Blackbox	6256	26.9	26.9	89.9
	4 Severn andWessex	2195	9.4	9.4	99.3
	5 Paper hand entry	112	.5	.5	99.8
	6 PMETB Web - not submitted	42	.2	.2	100.0
	Total	23267	100.0	100.0	



Section H - Items 2 to 6. Overall satisfaction rating for the post



Section C - About your supervision. Mean of the 5 items. Score ranges from 0 to 100, with higher scores indicating better supervision

***Appendix 2*: Differences in mean satisfaction scores and mean supervision scoresbetween Specialty Groups.**

a) Satisfaction.

Descriptives

satisfaction Section H - Items 2 to 6. Overall satisfaction rating for the post

	N	Mean	Std. Deviation	Std. Error	95% Confidence Interval for Mean		Minimum	Maximum
					Lower Bound	Upper Bound		
1	391	75.21	15.028	.760	73.72	76.71	26	100
2 Anaesthetics	2512	80.44	14.305	.285	79.88	81.00	20	100
3 Emergency Medicine	623	78.00	14.445	.579	76.86	79.14	29	100
4 General Practice	4038	82.31	14.944	.235	81.85	82.77	16	100
5 Medicine	4805	78.04	14.903	.215	77.62	78.47	16	100
6 Obstetrics and Gynaecology	1258	75.95	14.972	.422	75.12	76.78	16	100
7 Occupational Medicine	37	79.08	16.915	2.781	73.44	84.72	25	100
8 Ophthalmology	434	82.89	14.495	.696	81.52	84.26	16	100
9 Paediatrics and Child Health	1770	79.15	13.895	.330	78.50	79.80	20	100
10 Pathology	517	81.51	14.372	.632	80.27	82.76	36	100
11 Psychiatry	2037	82.60	13.708	.304	82.00	83.19	21	100
12 Public Health	98	81.19	15.114	1.527	78.16	84.22	33	100
13 Radiology	951	79.70	13.923	.451	78.81	80.58	16	100
14 Surgery	3796	77.13	16.909	.274	76.59	77.67	16	100
Total	23267	79.46	15.125	.099	79.27	79.66	16	100

ANOVA

satisfaction Section H - Items 2 to 6. Overall satisfaction rating for the post

	Sum of Squares	df	Mean Square	F	Sig.
Between Groups	117207.9	13	9015.994	40.275	.000
Within Groups	5205433	23253	223.861		
Total	5322640	23266			

satisfaction Section H - Items 2 to 6. Overall satisfaction rating for the post

Ryan-Einot-Gabriel-Welsch Range[a]

SpecGroupIntendedARC	N	Subset for alpha = .05				
		1	2	3	4	5
1	391	75.21				
6 Obstetrics and Gynaecology	1258	75.95				
14 Surgery	3796	77.13	77.13			
3 Emergency Medicine	623	78.00	78.00	78.00		
5 Medicine	4805		78.04	78.04		
7 Occupational Medicine	37		79.08	79.08	79.08	
9 Paediatrics and Child Health	1770			79.15	79.15	
13 Radiology	951			79.70	79.70	
2 Anaesthetics	2512				80.44	
12 Public Health	98				81.19	81.19
10 Pathology	517				81.51	81.51
4 General Practice	4038					82.31
11 Psychiatry	2037					82.60
8 Ophthalmology	434					82.89
Sig.		.151	.127	.285	.237	.950

Means for groups in homogeneous subsets are displayed.

a. Critical values are not monotonic for these data. Substitutions have been made to ensure monotonicity. Type I error is therefore smaller.

b) Supervision.

Descriptives

supervision Section C - About your supervision. Mean of the 5 items. Score ranges from 0 to 100, with higher scores indicating better supervision

	N	Mean	Std. Deviation	Std. Error	95% Confidence Interval for Mean		Minimum	Maximum
					Lower Bound	Upper Bound		
1	391	79.15	15.384	.778	77.62	80.68	15	100
2 Anaesthetics	2512	85.25	11.985	.239	84.78	85.72	5	100
3 Emergency Medicine	623	81.77	13.649	.547	80.69	82.84	25	100
4 General Practice	4038	83.29	13.877	.218	82.86	83.72	6	100
5 Medicine	4805	80.37	13.618	.196	79.98	80.75	5	100
6 Obstetrics and Gynaecology	1258	82.58	13.022	.367	81.86	83.30	19	100
7 Occupational Medicine	37	85.17	19.151	3.148	78.78	91.55	13	100
8 Ophthalmology	434	81.03	13.873	.666	79.73	82.34	30	100
9 Paediatrics and Child Health	1770	85.51	11.632	.276	84.96	86.05	20	100
10 Pathology	517	89.03	11.344	.499	88.05	90.01	31	100
11 Psychiatry	2037	87.80	11.186	.248	87.31	88.28	20	100
12 Public Health	98	85.46	12.981	1.311	82.86	88.06	38	100
13 Radiology	951	80.91	13.852	.449	80.03	81.80	5	100
14 Surgery	3796	78.82	15.269	.248	78.33	79.30	5	100
Total	23267	82.58	13.736	.090	82.41	82.76	5	100

ANOVA

supervision Section C - About your supervision. Mean of the 5 items. Score ranges from 0 to 100, with higher scores indicating better supervision

	Sum of Squares	df	Mean Square	F	Sig.
Between Groups	199093.8	13	15314.910	84.977	.000
Within Groups	4190761	23253	180.225		
Total	4389855	23266			

supervision Section C - About your supervision. Mean of the 5 items. Score ranges from 0 to 100, with higher scores indicating better supervision

Ryan-Einot-Gabriel-Welsch Range[a]

SpecGroupIntendedARC	N	Subset for alpha = .05					
		1	2	3	4	5	6
14 Surgery	3796	78.82					
1	391	79.15	79.15				
5 Medicine	4805		80.37	80.37			
13 Radiology	951		80.91	80.91			
8 Ophthalmology	434		81.03	81.03	81.03		
3 Emergency Medicine	623			81.77	81.77		
6 Obstetrics and Gynaecology	1258				82.58		
4 General Practice	4038				83.29		
7 Occupational Medicine	37				85.17	85.17	
2 Anaesthetics	2512					85.25	
12 Public Health	98					85.46	
9 Paediatrics and Child Health	1770					85.51	
11 Psychiatry	2037						87.80
10 Pathology	517						89.03
Sig.		1.000	.544	.642	.248	1.000	.655

Means for groups in homogeneous subsets are displayed.

a. Critical values are not monotonic for these data. Substitutions have been made to ensure monotonicity. Type I error is therefore smaller.

***Appendix 3*: The plotting of results, confidence intervals, multiple comparisons, and the Bonferroni correction**.

It is common in multilevel modelling to plot a graph in which cases (providers, deaneries or whatever) are plotted from smallest residual ("worst") to largest residual ("best"), as with figures 3a here. That is not problematic. What is problematic is that those graphs often also show confidence intervals around each point, and that allows a visual comparison (as with figure 3b). However, as in analysis of variance, there are many technical problems once more than a single comparison of two cases is being made.

A single paired comparison. Goldstein (7) points out (p.36) that if, *a priori*, one wishes only to compare two cases (schools in Goldstein's case), then it is straightforward to construct a confidence interval and make the comparison at, say, the 5% level. However, that is typically not the case, and here it would, for instance, only be the case if we wished, say, to compare two providers *and those two providers alone*.

Many individuals each making a single paired comparison. Goldstein and Healy (8) considered a more common situation in which, say, a graph is presented with many schools on it, and a pair of teachers wished to ask whether their own two schools were significantly different. In that case the best procedure is to plot confidence intervals $\pm$ 1.39 SEs, in which case if the bars do not overlap then those *two* schools do not differ significantly. The method is not however valid if the teachers then start to compare further pairs of schools.

Comparison of triplets. As Goldstein (7) points out, wider confidence intervals, by a factor of about 25% or so, are required, even if three schools are being compared rather than two (since three comparisons are being made implicitly).

Multiple comparisons. It is more than possible that some individuals looking at a typical plot of effects will wish to ask questions such as "Is the worst case significantly worse than the best case". This comparison is *a posteriori* since the best and the worst were not known in advance of the analysis. In such situations, Goldstein and Healy (8) merely say, "For these situations a suitable multiple-comparisons procedure will be required" (p.177), without commenting further. However there are very many multiple-comparison procedures, all subtly different from one another, some of which are much more conservative than others.

Comparison with the population average. The *MLwiN* manual (9) considers the situation in which the key question of interest is whether or not individual units differ from the population average. In that case they set the comparative standard deviation[4] at 1.96, and plot the means in order, looking for those which do not overlap with the overall mean. They make no comment on possible problems of multiple testing. I am presenting results here in a similar way, and any judgements made are made at the risk of those making them.

Comparative standard errors and diagnostic standard errors. Goldstein makes clear (p.25) that diagnostic standard errors are standardised, and should only be used for assessing the normality or otherwise of distributions. Comparative standard errors are used for comparing means with other values, be they the overall mean or other units in the analysis. See Appendix 2.2 of Goldstein for the difference in the computational formulae.

[4] Although the manual refers to "comparative standard deviation (SD)" (p.38), Goldstein refers to "comparative standard errors" on both p.25 and 36. I would have thought that standard error is the correct term.

(1) Benjamini Y, Drai D, Elmer G, Kafkaki N, Golani I. Controlling the false discovery rate in behavior genetics research. BBR 2001;125:279-84.

(2) Benjamini Y, Hochberg Y. Controlling the false discovery rate: A practical and powerful approach to multiple testing. Journal of the Royal Statistical Society, Series B 1995;57:289-300.

(3) Benjamini Y, Liu W. A step-down multiple hypotheses testing procedure that controls the false discovery rate under independence. Journal of Statistical Planning and Inference 1999;82:163-70.

(4) Benjamini Y, Liu W. A distribution-free multiple test procedure that controls the false discovery rate. Tel Aviv University: http://www.math.tau.ac.il/%7Eybenja/BL2.pdf; 1999.

(5) Afshartous D, Wolf M. Avoiding data snooping in multilevel and mixed effects models. Journal of the Royal Statistical Society, Series B 2007;in press.

(6) Poloniecki J. Half of all doctors are below average. Brit Med J 1998;316:1734-6.

(7) Goldstein H. Multilevel statistical models. 2 ed. London: Arnold; 1995.

(8) Goldstein H, Healy MJR. The Graphical Presentation of a Collection of Means. Journal of the Royal Statistical Society Series A (Statistics in Society) 1995;158(1):175-7.

(9) Rasbash J, Steele F, Browne W, Prosser B. A user's guide to *MLwiN* version 2.0. Bristol: Centre for Multilevel Modelling; 2005.

Figure 1

Simple model, multilevel variances only

a.

$$\text{satisfac}_i \sim \text{N}(XB,\ \Omega)$$

$$\text{satisfac}_i = \beta_{0i}\text{cons}_i$$

$$\beta_{0i} = 80.566(0.326) + u^{(4)}_{0,deanery(i)} + u^{(3)}_{0,provider(i)} + u^{(2)}_{0,specWITHINprov(i)} + e_{0i}$$

$$\begin{bmatrix} u^{(4)}_{0,deanery(i)} \end{bmatrix} \sim \text{N}(0,\ \Omega_u^{(4)}) \ :\ \Omega_u^{(4)} = \begin{bmatrix} 1.006(0.647) \end{bmatrix}$$

$$\begin{bmatrix} u^{(3)}_{0,provider(i)} \end{bmatrix} \sim \text{N}(0,\ \Omega_u^{(3)}) \ :\ \Omega_u^{(3)} = \begin{bmatrix} 8.150(1.499) \end{bmatrix}$$

$$\begin{bmatrix} u^{(2)}_{0,specWITHINprov(i)} \end{bmatrix} \sim \text{N}(0,\ \Omega_u^{(2)}) \ :\ \Omega_u^{(2)} = \begin{bmatrix} 29.020(1.878) \end{bmatrix}$$

$$\begin{bmatrix} e_{0i} \end{bmatrix} \sim \text{N}(0,\ \Omega_e) \ :\ \Omega_e = \begin{bmatrix} 196.794(1.912) \end{bmatrix}$$

*-2*loglikelihood(IGLS Deviance)* = 190918.600(23267 of 23267 cases in use)

Simple model, multilevel variances plus complex variation at level 1

b.

$$\text{satisfac}_i \sim \text{N}(XB,\ \Omega)$$

$$\text{satisfac}_i = \beta_{0i}\text{cons}_i + e_{2i}\text{satisfac2}_i$$

$$\beta_{0i} = 85.871(0.276) + u^{(4)}_{0,deanery(i)} + u^{(3)}_{0,provider(i)} + u^{(2)}_{0,specWITHINprov(i)} + e_{0i}$$

$$\begin{bmatrix} u^{(4)}_{0,deanery(i)} \end{bmatrix} \sim \text{N}(0,\ \Omega_u^{(4)}) \ :\ \Omega_u^{(4)} = \begin{bmatrix} 0.637(0.461) \end{bmatrix}$$

$$\begin{bmatrix} u^{(3)}_{0,provider(i)} \end{bmatrix} \sim \text{N}(0,\ \Omega_u^{(3)}) \ :\ \Omega_u^{(3)} = \begin{bmatrix} 7.547(1.251) \end{bmatrix}$$

$$\begin{bmatrix} u^{(2)}_{0,specWITHINprov(i)} \end{bmatrix} \sim \text{N}(0,\ \Omega_u^{(2)}) \ :\ \Omega_u^{(2)} = \begin{bmatrix} 20.801(1.421) \end{bmatrix}$$

$$\begin{bmatrix} e_{0i} \\ e_{2i} \end{bmatrix} \sim \text{N}(0,\ \Omega_e) \ :\ \Omega_e = \begin{bmatrix} 619.057(11.371) & \\ -2.692(0.062) & 0 \end{bmatrix}$$

*-2*loglikelihood(IGLS Deviance)* = 187985.200(23267 of 23267 cases in use)

Figure 2

Final model with multiple fixed effects at trainee level, multilevel variances and complex variation at level 1

$$\text{satisfac}_i \sim \text{N}(XB,\ \Omega)$$

$$\begin{aligned}\text{satisfac}_i = {} & \beta_{0i}\text{cons}_i + \text{-0.703(0.601)specIty\_3}_i + \text{2.726(1.076)specIty\_4}_i + \\ & \text{-3.505(0.481)specIty\_5}_i + \text{-5.310(0.574)specIty\_6}_i + \\ & \text{-3.206(2.331)specIty\_7}_i + \text{0.845(0.797)specIty\_8}_i + \\ & \text{-2.393(0.545)specIty\_9}_i + \text{-0.201(0.793)specIty\_10}_i + \\ & \text{1.288(0.574)specIty\_11}_i + \text{-0.458(1.512)specIty\_12}_i + \\ & \text{-1.455(0.719)specIty\_13}_i + \text{-3.095(0.499)specIty\_14}_i + \\ & \text{-0.121(0.178)sex\_2}_i + \text{0.637(0.385)route\_2}_i + \\ & \text{-0.323(0.457)route\_3}_i + \text{-0.015(0.789)route\_4}_i + \\ & \text{1.335(1.305)route\_5}_i + \text{-1.177(2.015)route\_6}_i + \\ & \text{0.025(0.023)yrqual}_i + \text{2.163(0.230)grade\_2}_i + \\ & \text{0.395(0.390)grade\_3}_i + \text{4.303(1.172)grade\_4}_i + \\ & \text{5.519(1.021)grade\_5}_i + \text{0.799(0.123)timepost}_i + e_{1i}\text{satisfac2}_i\end{aligned}$$

$$\beta_{0i} = \text{32.907(46.294)} + u^{(4)}_{0,deanery(i)} + u^{(3)}_{0,provider(i)} + u^{(2)}_{0,specWITHINprov(i)} + e_{0i}$$

$$\begin{bmatrix} u^{(4)}_{0,deanery(i)} \end{bmatrix} \sim \text{N}(0,\ \Omega_u^{(4)}) \ :\ \Omega_u^{(4)} = \begin{bmatrix} \text{0.779(0.408)} \end{bmatrix}$$

$$\begin{bmatrix} u^{(3)}_{0,provider(i)} \end{bmatrix} \sim \text{N}(0,\ \Omega_u^{(3)}) \ :\ \Omega_u^{(3)} = \begin{bmatrix} \text{2.140(0.653)} \end{bmatrix}$$

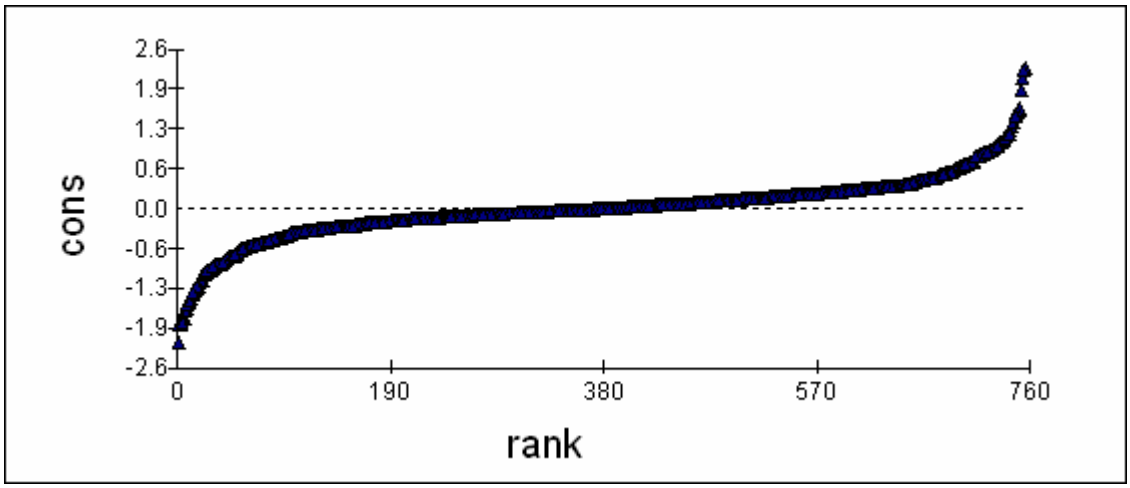
$$\begin{bmatrix} u^{(2)}_{0,specWITHINprov(i)} \end{bmatrix} \sim \text{N}(0,\ \Omega_u^{(2)}) \ :\ \Omega_u^{(2)} = \begin{bmatrix} \text{12.591(1.046)} \end{bmatrix}$$

$$\begin{bmatrix} e_{0i} \\ e_{1i} \end{bmatrix} \sim \text{N}(0,\ \Omega_e) \ :\ \Omega_e = \begin{bmatrix} \text{606.764(11.177)} & \\ \text{-2.625(0.061)} & 0 \end{bmatrix}$$

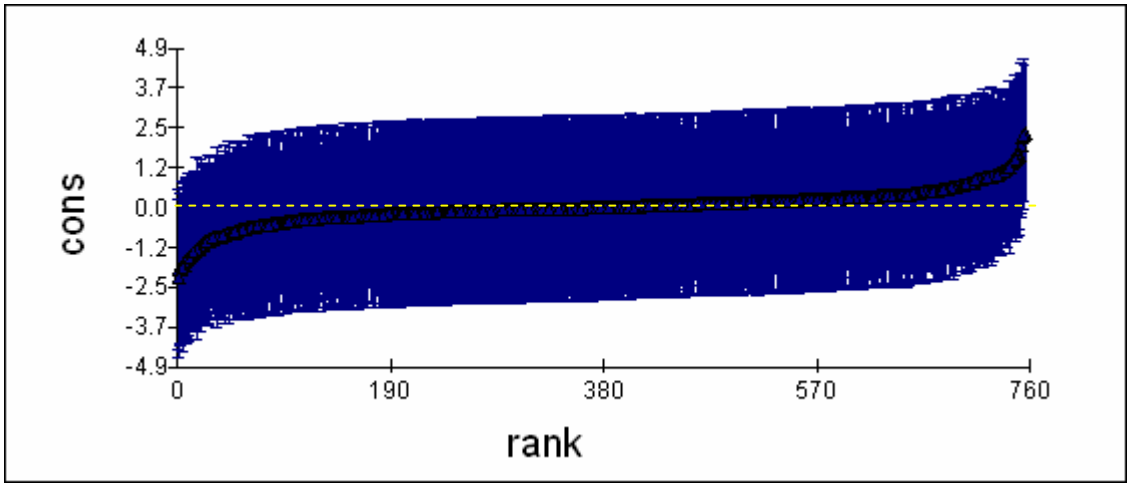
$$\textit{-2*loglikelihood(IGLS Deviance)} = \text{187258.700(23267 of 23267 cases in use)}$$

Figure 3 **Provider level effects (residuals)**

a.



b. Error bars show 95% confidence interval and are only for the purpose of comparing a single residual with the population mean



c.

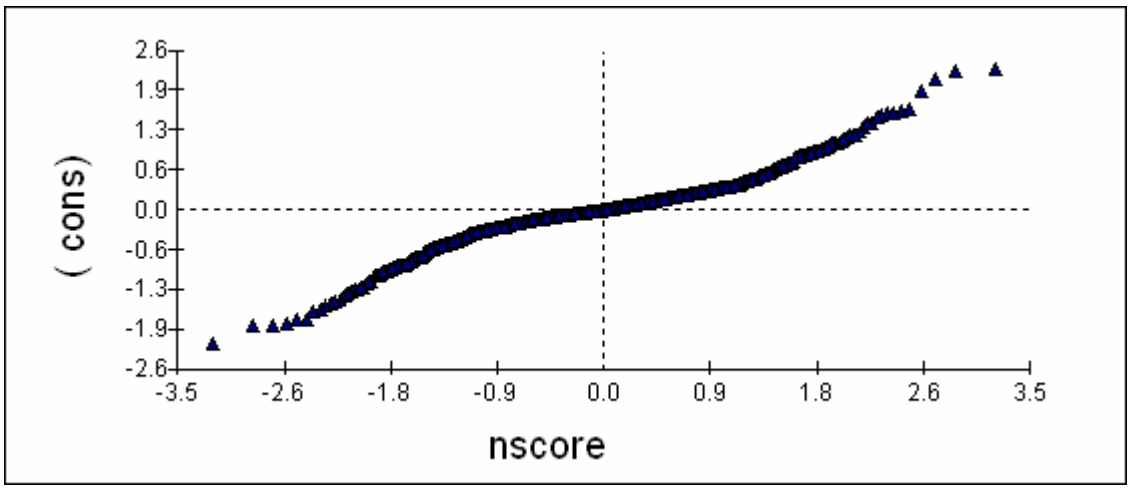
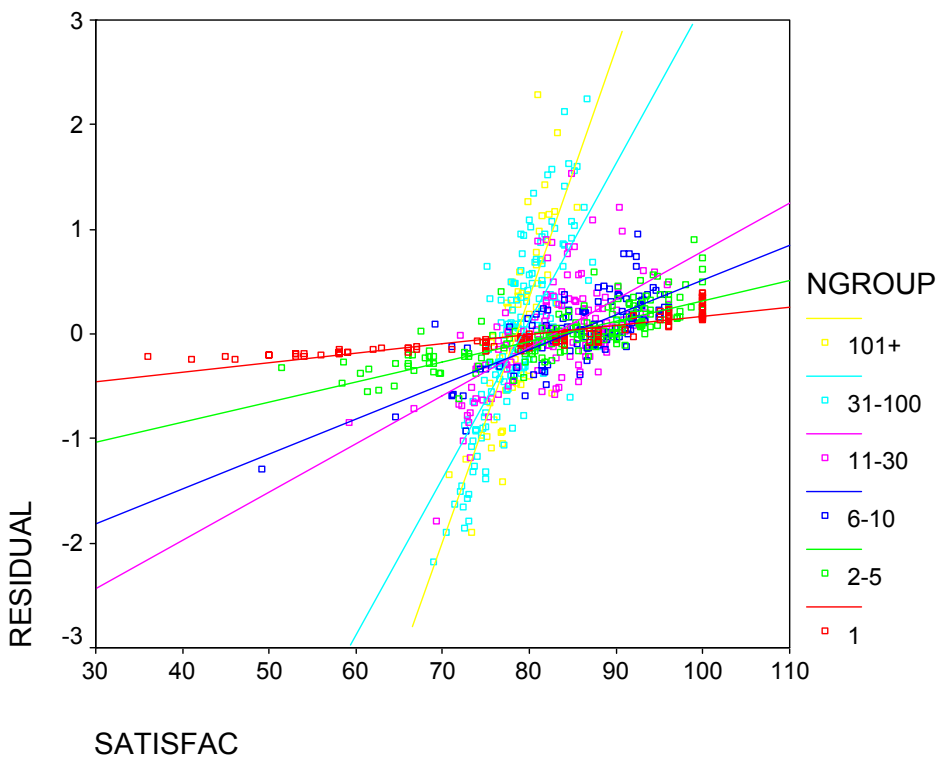


Figure 4 **Provider level effects (residuals)**

a.



b.

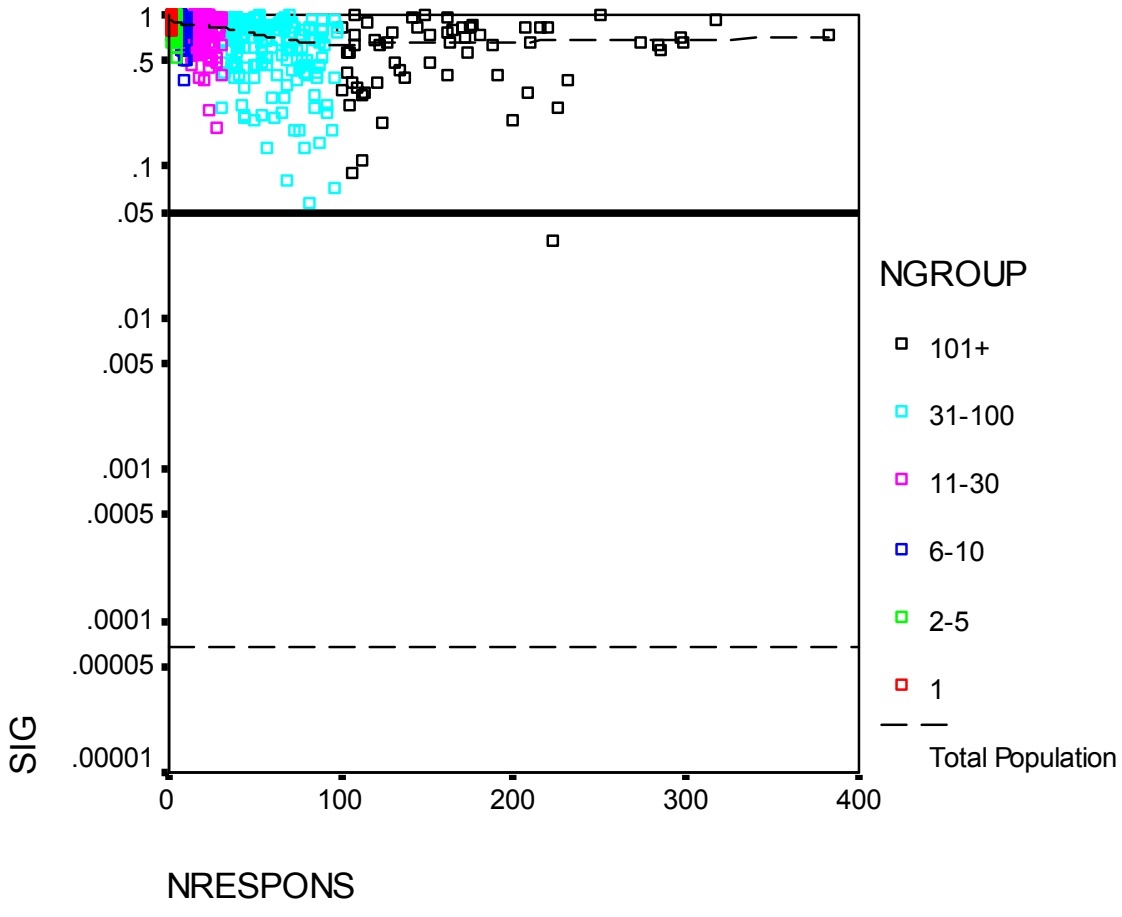
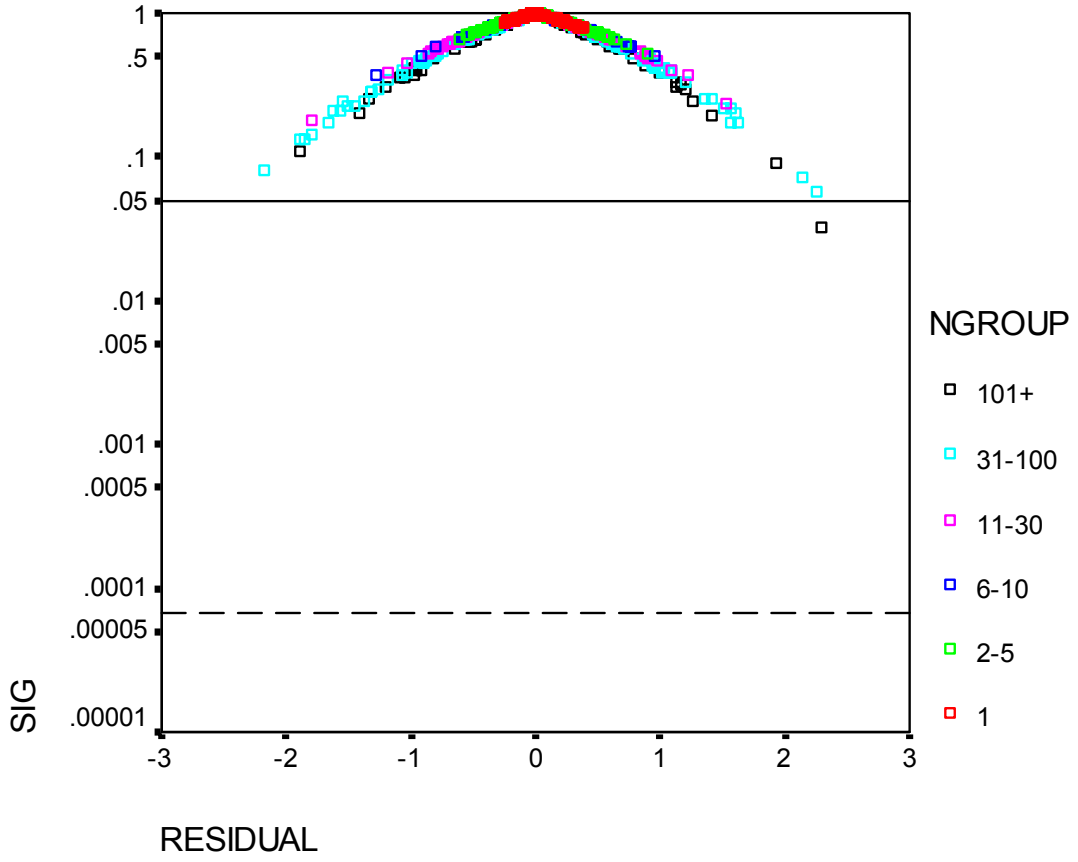


Figure 5 **Provider level effects (residuals)**

a.



b.

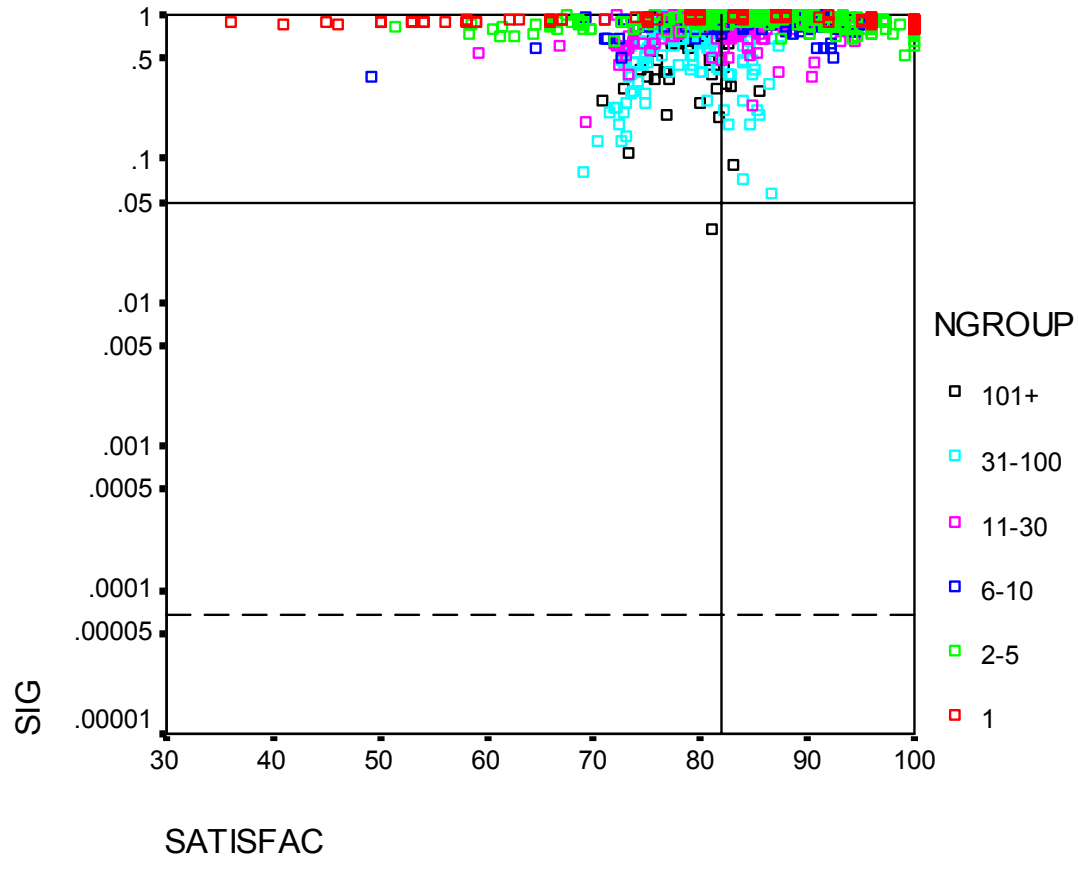
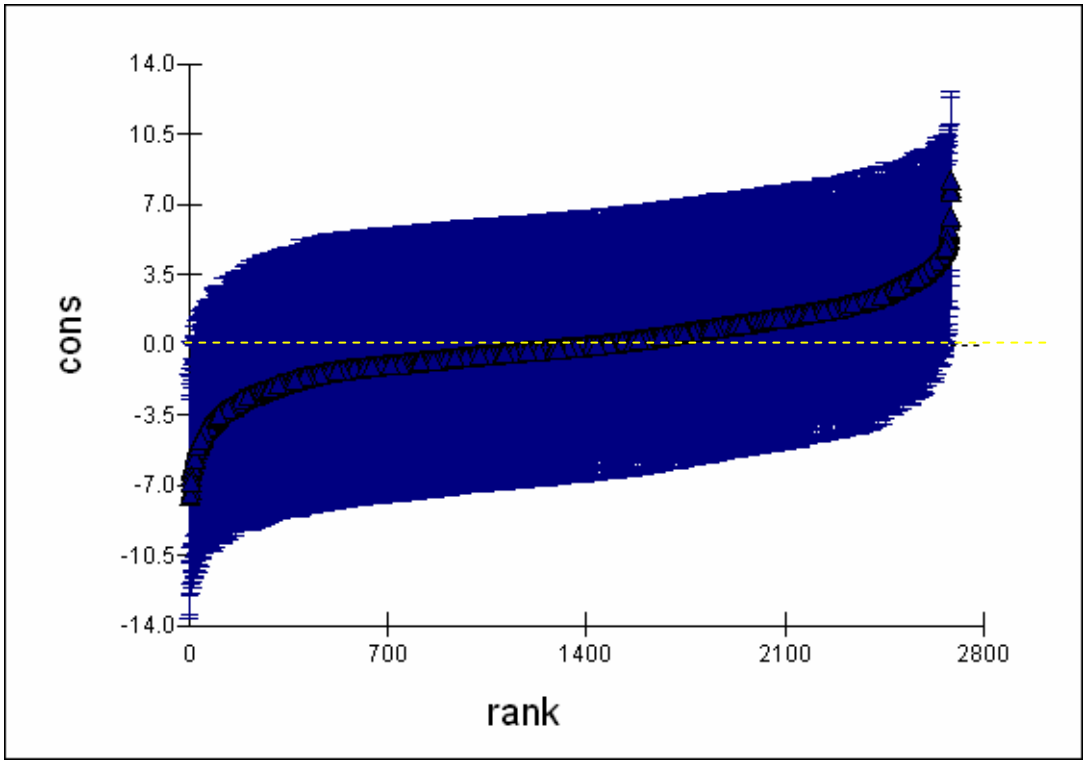


Figure 6 **Speciality within provider level effects (residuals)**

a. Error bars show 95% confidence interval and are only for the purpose of comparing a single residual with the population mean



b.

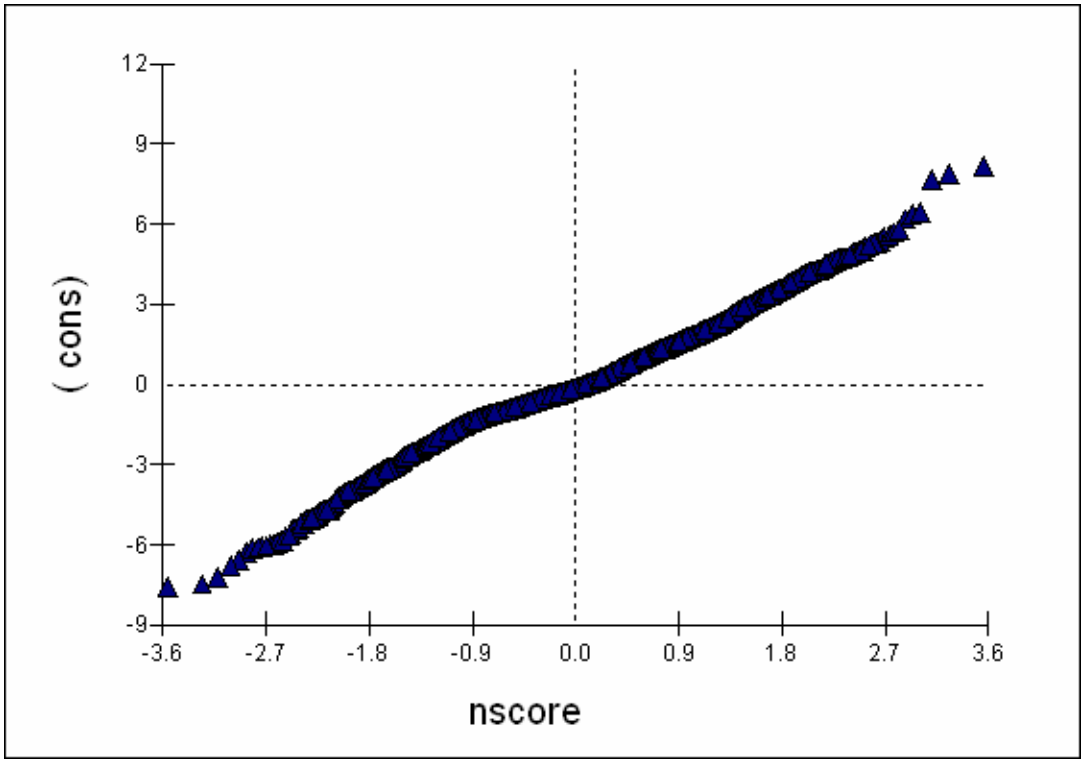
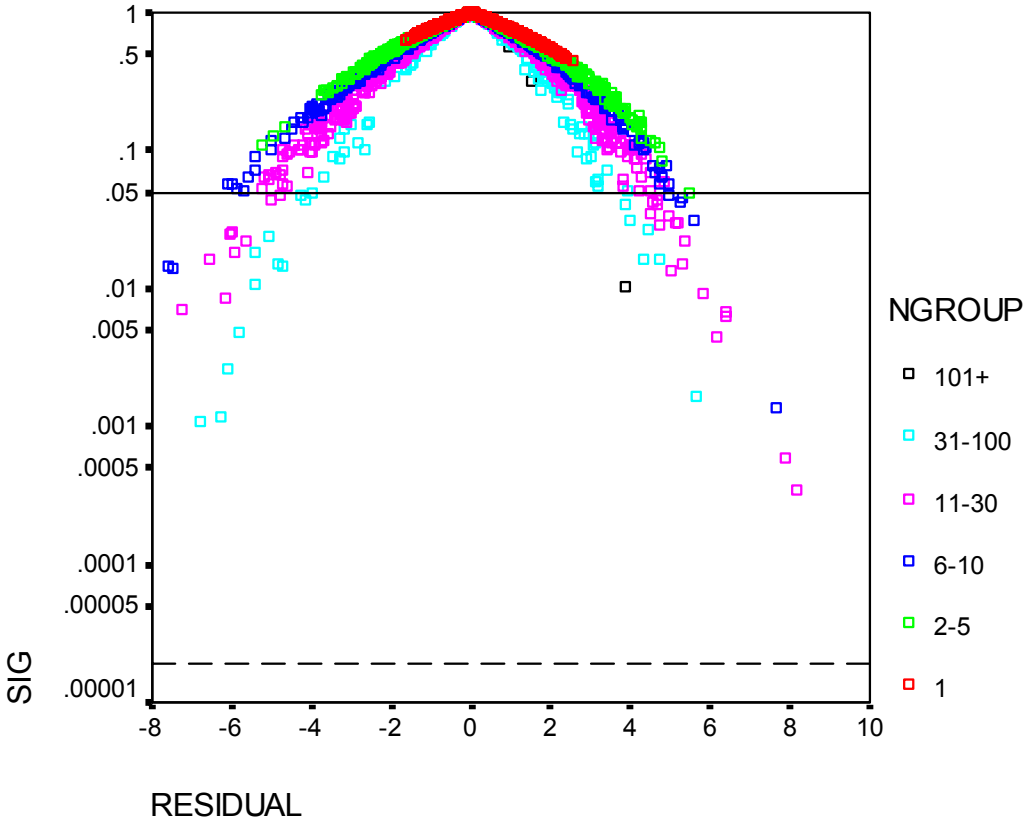


Figure 7

Speciality within trust level effects (residuals)

a.



b.

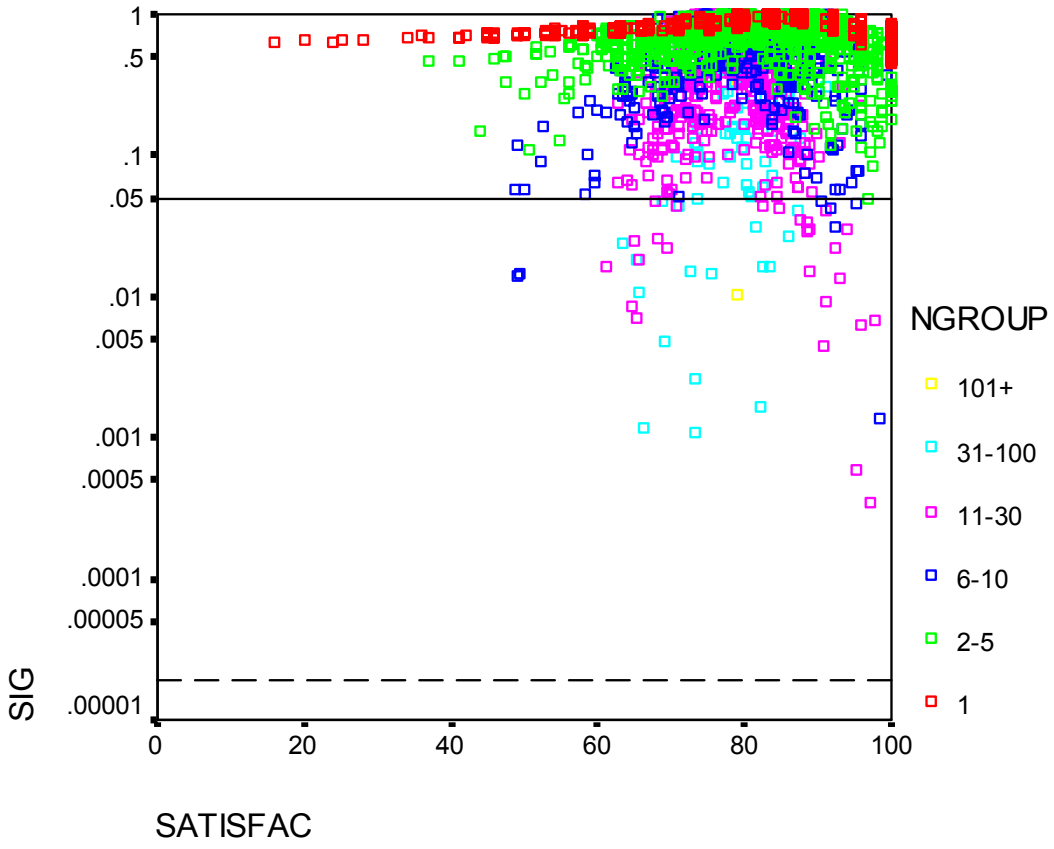


Figure 8

Final model with multiple fixed effects at trainee level, multilevel variances and complex variation at level 1

$$\text{supervis}_i \sim N(XB,\ \Omega)$$

$$\begin{aligned}\text{supervis}_i = {} & \beta_{0i}\text{cons}_i + 0.528(0.265)\text{route\_2}_i + -1.764(0.315)\text{route\_3}_i + \\ & 1.542(0.483)\text{route\_4}_i + 0.636(1.026)\text{route\_5}_i + \\ & -0.722(1.612)\text{route\_6}_i + -1.633(0.439)\text{speclty\_3}_i + \\ & -0.987(0.843)\text{speclty\_4}_i + -4.572(0.337)\text{speclty\_5}_i + \\ & -3.993(0.411)\text{speclty\_6}_i + 1.323(1.567)\text{speclty\_7}_i + \\ & -3.652(0.624)\text{speclty\_8}_i + 0.071(0.378)\text{speclty\_9}_i + \\ & 2.254(0.536)\text{speclty\_10}_i + 1.956(0.394)\text{speclty\_11}_i + \\ & -1.236(1.121)\text{speclty\_12}_i + -4.019(0.523)\text{speclty\_13}_i + \\ & -5.252(0.353)\text{speclty\_14}_i + -0.734(0.142)\text{sex\_2}_i + \\ & 3.612(0.163)\text{grade\_2}_i + -0.352(0.316)\text{grade\_3}_i + \\ & 3.854(0.939)\text{grade\_4}_i + 5.150(0.810)\text{grade\_5}_i + \\ & -0.114(0.097)\text{timepost}_i + e_{1i}\text{supervis2}_i\end{aligned}$$

$$\beta_{0i} = 89.855(0.472) + u^{(4)}_{0.\text{deanery}(i)} + u^{(3)}_{0.\text{provider}(i)} + u^{(2)}_{0.\text{specWITHIIIprov}(i)} + e_{0i}$$

$$\begin{bmatrix} u^{(4)}_{0.\text{deanery}(i)} \end{bmatrix} \sim N(0,\ \Omega_u^{(4)}) \ :\ \Omega_u^{(4)} = \begin{bmatrix} 0.175(0.127) \end{bmatrix}$$

$$\begin{bmatrix} u^{(3)}_{0.\text{provider}(i)} \end{bmatrix} \sim N(0,\ \Omega_u^{(3)}) \ :\ \Omega_u^{(3)} = \begin{bmatrix} 0.869(0.310) \end{bmatrix}$$

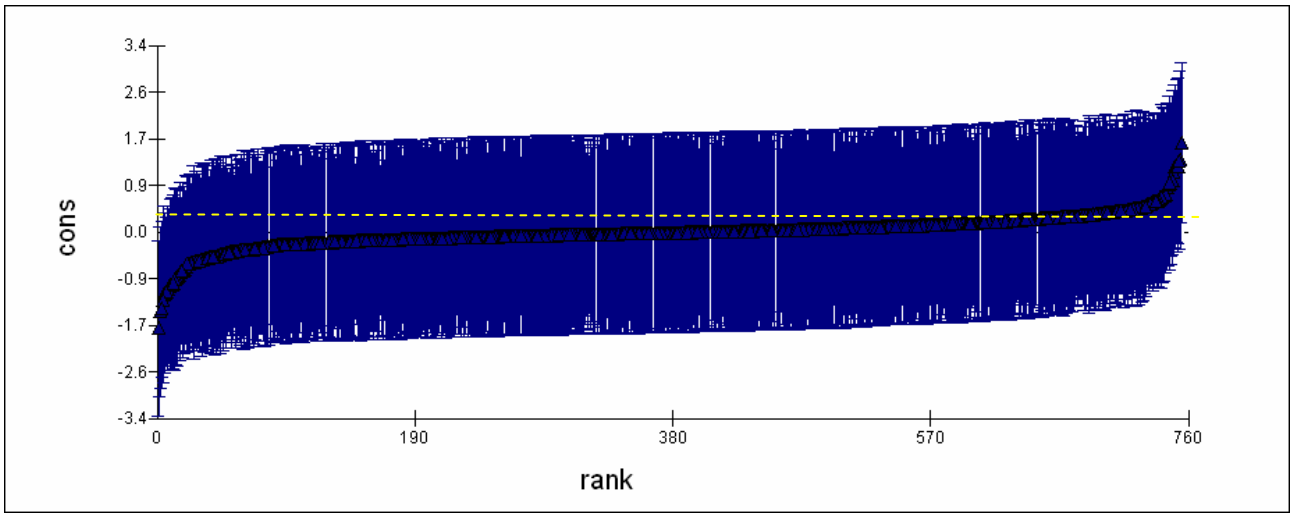
$$\begin{bmatrix} u^{(2)}_{0.\text{specWITHIIIprov}(i)} \end{bmatrix} \sim N(0,\ \Omega_u^{(2)}) \ :\ \Omega_u^{(2)} = \begin{bmatrix} 4.618(0.513) \end{bmatrix}$$

$$\begin{bmatrix} e_{0i} \\ e_{1i} \end{bmatrix} \sim N(0,\ \Omega_e) \ :\ \Omega_e = \begin{bmatrix} 632.664(10.068) & \\ -2.963(0.053) & 0 \end{bmatrix}$$

*-2*loglikelihood(IGLS Deviance)* = 178633.500(23267 of 23267 cases in use)

Figure 9 **Provider level effects (residuals)**

a. Error bars show 95% confidence interval and are only for the purpose of comparing a single residual with the population mean



Speciality within Provider level effects (residuals)

b. Error bars show 95% confidence interval and are only for the purpose of comparing a single residual with the population mean