

**NATIONAL STUDENT SURVEY OF TEACHING IN UK
UNIVERSITIES: DIMENSIONALITY, MULTILEVEL STRUCTURE,
AND DIFFERENTIATION AT THE LEVEL OF UNIVERSITY AND
DISCIPLINE: PRELIMINARY RESULTS**

Herbert W. Marsh and Jacqueline Cheng

Department of Education, Oxford University

28 August, 2008.

Author Note:

We would like to thank colleagues from the Higher Education Funding Council for England and the Higher Education Academy who provided support for this research, particularly Paul Ramsden, Mike Prosser, Matthew Watkins, Malgorzata Kulej, and Mark Gittoes. We would also like to acknowledge helpful suggestions at various stages of the research from Harvey Goldstein, John Richardson, Roger Brown, and Francesca Scalas. Requests for further information about this investigation should be sent to Professor Herbert W. Marsh, Department of Education, University of Oxford, 15 Norham Gardens, Oxford OX2 6PY UK; E-Mail: herb.marsh@education.ox.ac.uk.

Keywords: higher education; quality assurance; teaching effectiveness; multilevel analysis; factor analysis; university student satisfaction

TABLE OF CONTENTS

LIST OF TABLES	3
LIST OF FIGURES	4
EXECUTIVE SUMMARY	5
1 INTRODUCTION.....	10
1.1 Students Evaluations of Teaching Effectiveness	11
1.2 Relevant Australian Research.....	14
2 PROPOSED RESEARCH.....	17
3 NSS: BACKGROUND	20
3.1 Factor structure	20
3.2 Reliability.....	20
3.3 Ability to Differentiate Between Institutions and Discipline-within-university Groups21	
3.4 The Role of Individual Student, Course, and Institutional Characteristics.....	21
4 METHODS	22
4.1 Participants (and background characteristics)	22
4.2 Materials: NSS Survey Items.....	22
5 RESULTS	24
5.1 Factor Structure.....	24
Role of the Overall Satisfaction Rating	26
5.2 Multi-level Analyses: Appropriate Unit of Analysis	37
5.3 Differentiation as a Function of Error and Reliability	42
6 SUMMARY & DISCUSSION	50
7 REFERENCES.....	53
APPENDIX 1 - THE 22 NSS ITEMS BROKEN DOWN INTO 8 FACTORS.....	58
APPENDIX 2 – CATERPILLAR PLOTS OF UNIVERSITIES	59
APPENDIX 3 – CATERPILLAR PLOTS OF DISCIPLINES WITHIN UNIVERSITIES.....	63
APPENDIX 4 – FIXED EFFECTS ASSOCIATION WITH DUMMY VARIABLES OF EACH OF THE DISCIPLINE CLASSIFICATIONS USED IN 2005 AND 2006.....	67
APPENDIX 5 - FIXED EFFECTS ASSOCIATED WITH DUMMY VARIABLES OF EACH OF THE STUDENT CHARACTERISTICS	69
APPENDIX 6A - FREQUENCY DISTRIBUTION OF STUDENTS IN DIFFERENT “DISCIPLINE- WITHIN-UNIVERSITY” GROUPS FOR DIFFERENT DISCIPLINE CLASSIFICATIONS.....	71
APPENDIX 6 B - FREQUENCY DISTRIBUTION OF RELIABILITY ESTIMATES FOR THE AVERAGE OVERALL SATISFACTION RATING IN DIFFERENT “DISCIPLINE-WITHIN- UNIVERSITY” GROUPS FOR DIFFERENT DISCIPLINE CLASSIFICATIONS	73

LIST OF TABLES

Table 1. Explanatory variables considered

Table 2. Tests of the variance explained and statistical significance based on different numbers of factors

Table 3. Factor Pattern Loadings for Six- and Seven-factor Solutions (2005 Data)

Table 4. Summary of goodness of fit for all models

Table 5. Model 1: NSS 7 factors 22 items: 2005, 2006 data

Table 6. Model 2: NSS 8 factors 22 items: 2005, 2006 data

Table 7. Residual variance associated with university, discipline and student level of analysis when considering differences between disciplines within and between universities

LIST OF FIGURES

- Figure 1.** Comparison of Overall Ratings in 35 Australian universities ranked from highest to lowest
- Figure 2.** Scree Plot as a basis of determining the number of factors for 2005 and 2006 data
- Figure 3.** 7-factor and 8-factor structures with 1 higher-order factor of overall satisfaction.
- Figure 4.** Distribution of sample sizes across universities for 2005 (above) and 2006 (below) data.
- Figure 5.** Distribution of reliability estimates across universities for 2005 (above) and 2006 (below) data.
- Figure 6.** Scatter diagram of relations between ranking of universities based on the Overall Satisfaction rating for 2005 and 2006 (see Appendix 5 for the actual values for each university).

EXECUTIVE SUMMARY

The United Kingdom, Australia, and many other countries are seeking ways to improve the quality and enhance the accountability of its higher-education sector. Consistent with this aim the UK funding agencies and UK universities have cooperated in 2005 to collect standardised data that can be used “to provide the public and the higher education sector with comprehensive, comparable views of students about the quality of their education” (HEFCE, 2005, p. 2).

The National Student Survey (NSS) published its first results in 2005 on the Teaching Quality Information (TQI) website. These results are also readily available though published sources such as “*The Times Good University Guide 2007*” published in association with Price Waterhouse Coopers. Providing a strong argument for the appropriateness of comparing universities on a wide variety of different indicators, The Good University Guide uses the average response to 22 NSS items to measure overall “student satisfaction” – one of the criteria used to construct “league tables” comparing different UK universities.

The aims of the preliminary analyses were to evaluate the factor structure implicit in the design of the NSS responses as well as more complex multilevel analyses to explore the replicability of the structure and proportions of variance that can be explained by background variables at the level of the individual student, discipline (discipline groups within each university), and the university. The NSS was designed to measure six factors (Teaching, Assessment and Feedback, Academic Support, Organisation and Management, Learning Resources, Personal Development), plus an Overall Satisfaction item. Each factor was based on multiple items, a total of 22 items in all (see appendix A for a listing of the items and the factor each is designed to measure).

ANALYSES CONDUCTED

1. Factor analysis

Survey instruments are typically designed to measure specific aspects of the construct that is being surveyed – the factors. Each factor is based on responses to several items designed to measure that factor. For example, the NSS was designed to measure six factors based on responses to 22 items making up its factor structure. An important evaluation of the validity of responses to a survey instrument is whether the factors that it actually measures match those that it is designed to measure. Such tests are based on factor analysis.

2. Multilevel modelling

In higher education settings, students are grouped within the courses or disciplines that they take, which are grouped under the umbrella of the university that they attend. Individual characteristics and those associated with groups to which they belong are confounded because these groups are typically not established according to random assignment. A multilevel perspective is critical in that the same variable can have very

different effects at different levels of analysis and reliability at one level (e.g., the individual student) provides no evidence that responses are reliable at a different level (e.g., university or discipline-within-university levels).

The appropriate unit of analysis is a critical methodological issue in the evaluation of NSS responses. In particular, if the purpose of the NSS ratings is to differentiate between universities (i.e., a benchmarking exercise) then university is a critical unit of analysis. However, if it is also important to differentiate between discipline-within-university groups, then discipline should also be an important unit of analysis. Fortunately, appropriate multilevel modelling approaches allow the incorporation of more than one unit of analysis (e.g., students, disciplines, universities) within the same analysis. From a practical perspective, a multilevel approach allows researchers to pursue new questions about how effects vary from group to group and the characteristics of groups associated with this variation.

MAIN FINDINGS

1. Factor analysis identified the six factors that the NSS was designed to measure (Teaching; Assessment & Feedback; Support; Organisation; Resources; Personal Development; see Appendix 1), but also suggested that assessment and feedback should be considered as separate factors. The same factor structure was found for 2005 and 2006 responses. Further research is needed to establish practical significance of this distinction.
2. Factor analyses also demonstrated that it would be appropriate to summarise NSS ratings with a single overall satisfaction score. This could be either a weighted average of the specific factors and overall rating item (a higher-order factor score) *or* simply the overall rating item by itself. However, information about the specific factors is lost by doing this. As the specific factors are not equally important in terms of their contribution to the overall summary it is probably *not* appropriate to take a simple unweighted average of the scale scores or the items to obtain a single summary score.
3. There is not substantial variation between different universities in terms of overall satisfaction; differences between universities explain only about 2.5% of the variance (variance component based on multilevel analyses, controlling for student characteristics and discipline). However, because the number of students from each institution is so large, the differences between institutions are quite reliable (for all but a handful of universities with small numbers of respondents).
4. Consistent with the substantial reliability observed at each year considered separately, the rank ordering of universities based on the 2005 and on the 2006 results is highly stable. The correlation between rankings in 2005 and 2006 was 0.86, and mean overall satisfaction rating was nearly identical in each year.

5. There is substantially more variance explained in overall satisfaction by differences between discipline groupings within universities than by differences between universities.
 - Some of this difference is due to discipline differences that generalise across universities. Hence, these differences are not particularly useful for differentiating between universities in terms of particular disciplines (e.g., bench-marking each discipline across universities).
 - Discipline differences that generalise across universities complicate comparisons among differences at any one university. For example, averaged across all universities, Historical and Philosophical Studies tend to get higher ratings than other disciplines. Hence, some of the apparently higher ratings of Historical and Philosophical Studies at any given university may reflect differences inherent to the discipline that are not specific to a particular university.
 - Even after controlling for discipline differences that generalise across universities, up to 4.5% of individual student variance (depending on the discipline classification) is specific to discipline groupings within universities. However, because the number of students in each discipline-within-university group can be quite small – particularly when many discipline categories are considered, mean ratings for these groups can lack reliability.
6. There is need for an appropriate balance between the number of disciplines considered (different discipline classifications vary substantially in the number of disciplines considered) and the number of students within each discipline-within-university groups. Discipline-within-university groups explain more variance when the number of discipline categories is larger (i.e., the discipline classification becomes more specific). However, because the number of students in each discipline-within-university group can be quite small – particularly when many discipline categories are considered, mean ratings for these groups can lack reliability.
7. Student and institution background characteristics do not explain a lot of the variance in the satisfaction ratings. Controlling for student characteristics has little effect on overall satisfaction at either the university level or the discipline-within-university-group levels.

KEY RECOMMENDATIONS

1. It is appropriate to summarise NSS ratings with a single overall satisfaction score based on the overall summary rating or an appropriately weighted average of the 22 NSS items. Because the contributions of the different first-order factors to the higher-order factor differed substantially, the results argue against the use of an unweighted average of NSS responses (factors or items). Also, there were substantial amounts of variance in ratings in many of the first-order factors that were not explained by the single global satisfaction factor. However, further research is needed to explore the practical implications of focusing primarily on overall satisfaction rather than specific factors – particularly in relation to providing information to universities and

discipline-within-university groups that is useful for enhancing the quality of the undergraduate education experience.

2. Variance components at the university level were highly significant from a statistical perspective and highly reliable – due primarily to the very large sample sizes at nearly all universities. However, differences between universities explained only about 2.5% of the variance in individual student responses after controlling for discipline and student characteristics. Hence, there is much more variation in responses by students within each university than there is in responses between the different universities (i.e., there is substantial lack of agreement among students within each university in terms of their satisfaction with their overall educational experience). Nevertheless, because of the high reliability (due to large number of students from each university), the differences between universities were highly stable over the two years that we considered ($r = .86$ indicating high agreement in the ranking of universities in the two years).

Interpretation of these university differences in overall satisfaction results provides a dilemma. Differences between universities explain only a small amount of the variance in the NSS responses. Thus, only a relatively few universities are significantly above – or below – the average across all universities. The critical question is whether these small, but highly reliable, differences between universities are sufficiently large to help inform the choices of prospective students—a primary purpose of the NSS.

To more accurately evaluate the extent of differences between universities and the results of any one university, the mean ratings should be supplemented with error bars. These error bars show the range of probable error in results for each university and more accurately represent the nature of differences between universities.

3. NSS ratings should only be used with appropriate caution to compare ratings by different discipline-within-university groups (either different disciplines within the same university or the same discipline across universities). Thus, for example, discipline differences that generalise across universities (e.g., Historical and Philosophical Studies receives higher ratings across all universities) complicates ratings of Historical and Philosophical Studies within a given university. Hence, comparisons of how Historical and Philosophical Studies compares with other disciplines within a given university should also be based in part on Historical and Philosophical Studies ratings across universities.

Interpretation of differences due to disciplines should be qualified in relation to interpretations of probable error based on appropriate multilevel models. In particular, discipline-within-university differences are typically not reliable (due to small sample sizes) and this problem is made clear by the inclusion of error bars.

4. Generic discipline classifications designed to be applicable across all universities vary in terms of their appropriateness at any one university. More work is needed to get the

best balance between the number disciplines to be considered, their appropriateness across universities, and the number of students in each discipline-within-university group. The balance should maximise differences between groups (based on more disciplines) while still having enough students to provide reliable results (based on fewer disciplines).

1 INTRODUCTION

Universities throughout the world are undertaking benchmarking exercises in which they compare themselves to other universities on appropriate indices in order to establish their current levels of performance and to initiate continuous self-improvement. In order to pursue benchmarking exercises, there is a need for a comprehensive set of benchmark indicators that focus on outcomes, measure functional effectiveness rather than criteria that are easily countable, are systematically developed so as to have good construct validity as well as content (and “face”) validity, and to differentiate between universities (or academic units within universities) so as to provide appropriate standards as a basis of ascertaining excellence and continuous improvement.

The United Kingdom, Australia, and many other countries are seeking ways to improve the quality and enhance the accountability of its higher-education sector. Consistent with this aim, the UK funding agencies and UK universities have cooperated since 2005 to collect standardised data that can be used “to provide the public and the higher education sector with comprehensive, comparable views of students about the quality of their education” (HEFCE, 2005, p. 2). More specifically, the NSS was specified as having three aims:

- 1) To help inform the choices of prospective students;
- 2) To contribute to public accountability; and
- 3) To provide useful data to institutions to use in their enhancement activities.

The first data collection took place between January and April 2005. The NSS was repeated in 2006 and 2007, and it is anticipated that the NSS (or some version of it based on research and experience with the initial version) will continue to be used in the future.

Results of the 2005 were published on the Teaching Quality Information (TQI) website. These results are also readily available through published sources such as 2006 Edition of The Times “Good University Guide 2007” published in association with Price Waterhouse Coopers. Readers are told that the NSS “is an annual survey of final-year students asking about their experiences as a student. It is not a direct measure of quality, but indicates how satisfied student were with the experience they had” (p. 28). Providing a strong argument for the appropriateness of comparing universities on a wide variety of different indicators, The Times Good University Guide uses the average response to 22 NSS items to measure “student satisfaction” – one of the criteria used to construct “league tables” comparing different UK universities.

Although there is limited research on the use of university student ratings to evaluate universities as a whole, there is a large research literature on the use of students’ evaluations of teaching effectiveness to evaluate the effectiveness of individual teachers (Marsh, 1984, 1987, 2007; Marsh & Dunkin, 1997; Marsh & Roche, 1993; 1997).

Importantly, the NSS is designed to measure student satisfaction with their overall university experience – which includes, but is not limited to, teaching effectiveness. Nevertheless, much insight can be gained from student evaluation research that is relevant to the present investigation. Hence, we begin with a brief review of selected aspects of this research that are particularly relevant to our evaluation of the NSS.

1.1 Students Evaluations of Teaching Effectiveness

In higher education, there is a long history of research and much debate into the use of students' evaluations of teaching effectiveness (e.g., d'Apollonia & Abrami, 1997; Feldman, 1997, 1998; Greenwald & Gillmore, 1997; Marsh, 1984, 1987, 2007; Marsh & Roche, 1997, 2000; McKeachie, 1997). Effective teaching is a hypothetical construct for which there is no adequate single indicator. Hence, the validity of students' evaluations of teaching or of any other indicator of effective teaching must be demonstrated through a construct validation approach. Extensive reviews of this research (e.g., Abrami, d'Apollonia, & Cohen, 1990; Cashin, 1988; Cohen, 1980; Feldman, 1989a, 1989b, 1997, 1998; Marsh, 1984, 1987, 2007; Marsh & Dunkin, 1997, McKeachie, 1979, 1997a, 1997b) have consistently shown that, with careful attention to measurement and theoretical issues, students' evaluations of teaching are:

- Multidimensional;
- Reliable and stable;
- Primarily a function of the instructor who teaches a course rather than the course that is taught;
- Relatively valid against a variety of indicators of effective teaching;
- Relatively unaffected by a variety of variables hypothesized as potential biases, such as expected course grades, class size, workload and prior subject interest and;
- Demonstrably useful in improving teaching effectiveness when coupled with concrete enhancement strategies in specific areas that teachers target for improvement.

Emphasising the individual teacher (or class-average) as the appropriate unit of analysis for students' evaluations of teaching effectiveness, Marsh and Roche (1997; also see Marsh, 1987) stressed that analyses must be conducted at the appropriate unit of analysis in relation to the intended use of the ratings. This student evaluation research provides one model of an ongoing research program to evaluate the reliability, stability, factor structure, construct validity, potential biases, and usefulness for improving practice based on the NSS.

Unit of Analysis Problem

The appropriate unit of analysis is a critical methodological issue in student evaluation research that has particular relevance to our evaluation of NSS responses. Fortunately, however, there is a clear consensus in student evaluation research that the class-average or individual teacher is the appropriate unit of analysis rather than the

individual student (e.g., Marsh, 1987, 2007). Thus, support for the construct validity of student evaluation responses can only be demonstrated at the class-average level and the reliability of responses is most appropriately determined from studies of interrater agreement that assess error due to the lack of agreement among different students within the same course (see Gilmore, Kane, & Naccarato, 1978 for further discussion).

The correlation between responses by any two students in the same class (i.e., the single-rater reliability; Marsh, 1987) is typically in the .20s. However, the reliability of the **class-average** responses depends upon the extent of agreement among students within the same class *and* the number of students rating the class: .95 for 50 students, .90 for 25 students, .74 for 10 students, and .60 for five students. Given a sufficient number of students in any one class (or, perhaps, averaged across different classes taught by the same teacher if the number of student in any one class is less small) the reliability of class-average ratings is very good. Similarly, support for the construct validity of student evaluation responses must be demonstrated at the class-average level (e.g., relations with class-average achievement, teacher self-evaluations).

In trying to separate the effects of the teacher and the course, Marsh (1987; Marsh & Dunkin, 1997) reported that the correlation between overall teacher ratings of different instructors teaching the same course (i.e., a course effect) was -.05, whereas correlations for the same instructor in different courses (.61) and in two different offerings of the same course (.72) were much larger. These results support the validity of student evaluations as a measure of teacher effectiveness, but not as a measure of the course quality that is independent of the teacher. Marsh and Bailey (1993) further demonstrated that each teacher has a characteristic profile on the different evaluation factors (e.g., high on organization and low on enthusiasm) that was distinct from the profiles of other instructors and generalized across course offerings over a 13-year period.

Although there is some research suggesting discipline differences (e.g., a weak tendency for higher ratings in humanities and lower ratings in sciences; see Centra, 1993), these effects account for very little variance and there is ongoing debate about how these differences should be interpreted. Indeed, in many student evaluation programs, ratings for a given class are “normed” in relation to similar classes (similar in terms of student composition, level, and discipline), implying that such differences may not be important.

Hence, at least for the content of items typically considered in this student evaluation research (e.g., enthusiasm, learning/value, organization, rapport, group interaction, breadth of coverage, examinations), the appropriate unit of analysis is the individual teacher and not the individual student. Although it may be possible to construct an alternative set of items that would capture the quality of a course or program that was reasonably independent of the effects of specific teachers, there is little empirical support for this possibility in the student evaluation literature.

In their review of relevant literature on the unit of analysis problem, Marsh, Rowe and Martin (2002) also considered briefly the large body research into school effectiveness research in both elementary and secondary schools. Historically, this

research has focused on standardized test scores as the only outcome measures and sometimes failed to take into account pre-existing differences in students enrolled in different schools. More recently, research in this area is based on multilevel models and sophisticated “value added” models. In contrast to earlier research that did not account for the inherent hierarchical structure of the data, this more recent research has clearly demonstrated that effective schools are primarily a function of effective teachers within these schools. Once class/teacher effects have been taken into account, there is little residual variance at the school level. Furthermore, even in the most effective schools there is substantial variation at the class/teacher level. Indeed, Monk (1992) cites a number of studies in support of the observation that: “One of the recurring and most compelling findings within the corpus of production function research is the demonstration that how much a student learns depends on the identity of the instructor to which that student is assigned” (p. 320). Whereas this school effectiveness research comes from a very different perspective than research on students’ evaluations of university teaching as reviewed earlier, both research literatures lead to a similar conclusion that the individual teacher is the most important unit of analysis in assessing the quality of education.

These concerns about the appropriate unit of analysis are particularly relevant for the present investigation of NSS responses in which our focus is on the overall undergraduate experience at the broad level of the university and discipline-within-university groups rather than the effectiveness of individual teachers. Extrapolations from student evaluation of teaching research and school effectiveness research suggest that there might not be substantial variation associated with university or with discipline-within-university groups. Hence, the unit of analysis issue is one of the critical complexities in the appropriate analysis of NSS responses and a methodological focus on multilevel modelling is an important component in the evaluation of these issues.

Students’ Evaluations Can Lead to Improved Teaching

One potential purpose of NSS responses is to provide informative feedback that will lead to the improvement of undergraduate programmes. There is clear evidence that feedback from students’ evaluations of teaching, coupled with appropriate consultation, can lead to improved teaching effectiveness (see reviews by Cohen, 1980; L’Hommedieu, Menges & Brinko, 1990; Marsh, 1987; Marsh & Dunkin, 1997; Marsh & Roche, 1993). For example, in a study by Marsh and Roche (1993), randomly assigned intervention- and control-group- teachers completed self-evaluations and were evaluated by students before and after the intervention. An essential component of the intervention was a set of teaching strategy booklets – one for each factor on the student evaluation instrument. Teachers selected the factor to be targeted in their individually structured intervention and then selected the most appropriate strategies from a book of strategies relevant to that factor.

Marsh and Roche (1993) showed that intervention teachers improved significantly more than control group teachers. Furthermore, for the intervention group (compared to control group), targeted dimensions improved substantially more than non-targeted

dimensions. The study demonstrated that feedback from students' evaluations of teaching and consultation are an effective means of improving teaching effectiveness. It is important to note that this intervention can only be conducted with a well-designed, multidimensional instrument and that the specificity of the intervention effects to the targeted dimensions further supports the construct validity of multidimensional students' evaluations of teachings.

The lessons from this research are that: the feedback from student ratings needs to be specific to each teacher; teachers may need concrete strategies about how to improve their teaching; and this feedback may need to be complemented by a trained consultant. Even when teachers are motivated to improve their teaching and have feedback about their strengths and weaknesses, they still need professional assistance on how to actually improve their teaching. This need for consultation and concrete strategies is not surprising in that university academics typically receive little training in how to be effective teachers compared to training in being effective researchers. Although clearly beyond the scope of the present investigation, there is need to determine the extent to which the NSS provides feedback to universities or to departments within universities that actually leads to improved educational quality.

1.2 Relevant Australian Research

Students' evaluations of teaching effectiveness have typically focused on the efforts of individual teachers (or classes) so that there is limited basis for making comparisons across universities (or across similar disciplines from different universities). Although there is little research into the systematic use of student surveys to compare the quality of teaching programs across large numbers of different universities, perhaps the most relevant is that based on the Australian Course Evaluation Questionnaire (CEQ) that has a similar rationale to the NSS used in the UK and the Australian Postgraduate Research Experience Questionnaire (PREQ).

In 1991, the Australian government commissioned trials of the Course Experience Questionnaire (CEQ) in order to monitor the quality of students' university experiences. The CEQ is now routinely completed by graduates from all Australian universities within a few months of graduation. The responses assess characteristics of good teaching and effective learning such as enthusiasm, feedback, clarity of explanations, the establishment of clear goals and standards, the development of generic skills, the appropriateness of the workload and assessment, and an emphasis on student independence (Ainley & Long, 1994; Johnson, 1998; Ramsden, 1991).

Like the NSS, the intent of the CEQ was to provide an overall perspective of student experience and the results of this exercise are broadly available, for example, through "The Good Universities Guide to Australian Universities" used by potential students to select universities. In the McKinnon, Walker, & Davis et al. (2000) Benchmarking exercise, the CEQ is specifically recommended (Benchmark 6.10) to monitor student ratings of their experience of teaching, goals and standards, assessment practices, workload, generic skills and overall satisfaction and has been validated in a

non-Australian setting (Byrne & Flood, 2003). Subsequently, studies using the CEQ have found positive correlations between positive student approaches to learning and course experience (Diseth, Pallesen, Hovland, & Larsen, 2006), perceptions of a fair learning environment (Lizzio, Wilson, & Hadaway, 2007) and has been used to evaluate students' experiences with new, implemented courses delivered in medical settings (Bligh, Lloyd-Jones, & Smith, 2000) and web-based settings (Richardson, 2006).

Based on the CEQ, the Australian government commissioned the development and evaluation of the Postgraduate Research Experience Questionnaire (PREQ) to provide a multidimensional measure of the experience of postgraduate research students as part of a large-scale national benchmarking exercise for Australian universities. The unit of analysis is a critical issue in this research. Marsh, Rowe and Martin (2002) argued – based on reviews of research in the areas of students' evaluations of university teaching, teacher/school effectiveness, and teacher improvement – that the most important unit of analysis was the individual supervisor. However, the intended focus of PREQ was on the overall postgraduate experience at the broad level of the university, and discipline-within-university groups, rather than the effectiveness of individual supervisors. Indeed, students were specifically asked not to name their supervisor, and some of the factors focused on departmental or university level issues. Marsh et al. evaluated the PREQ based on responses from 32 Australian and New Zealand Universities.

At the level of the individual student, responses had reasonable psychometric properties (factor structure and internal consistency estimates of reliability). Consistent with the potential use of these instruments to benchmark the quality of supervision across all Australian universities, Marsh et al. evaluated the extent to which responses reliably differentiated between universities, academic disciplines, and discipline-within-university groups. Based on two-level (individual student, university) and three-level (individual student, discipline, university) multilevel models, the responses failed to differentiate among universities, or among discipline-within-university groups.

Although there were small differences between disciplines on a few PREQ scales, even they were consistent across different universities. Thus, for example, the largest differences were for the PREQ Infrastructure scale (e.g., technical/financial support, computers). Whereas humanities received the lowest ratings on this scale, this tended to be the case across all universities. Hence these discipline differences appeared to be inherent to the discipline and were not a function of the supervisors, administration, policies and support in humanity programs in particular universities.

The results demonstrate that PREQ responses that are adequately reliable at one level (individual student) may have little or no reliability at another level (university). These results were subsequently replicated with the next wave of data collected the following year with a new cohort of PhD students. In particular, results in Figure 1 show that the range of probable error (the 95% confidence interval about the mean rating for each university) for each of the 35 Australian universities contained the mean averaged across all the universities. No one university differed significantly from the average score

across all universities. On this basis, Marsh et al. concluded that PREQ responses should not be used to benchmark Australian universities or discipline-within-university groups.

The most salient finding of this study was that PREQ ratings did not vary systematically between universities, or between discipline-within-university groups. This has critically important methodological and substantive implications for the potential usefulness of the PREQ ratings. Because there was no significant variation at the university level, it follows that the PREQ ratings were completely unreliable for distinguishing between universities. This clearly demonstrates why it is important to evaluate the reliability of responses to a survey instrument in relation to a particular application and the level of analysis that is appropriate to this application.

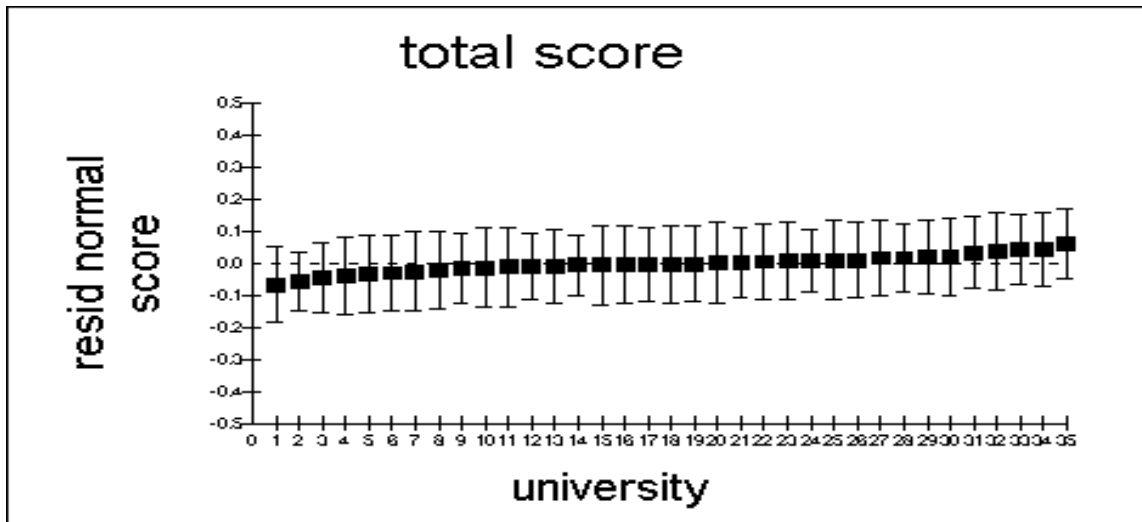


Figure 1. Comparison of Overall Ratings in 35 Australian universities ranked from highest to lowest. Results show that the error bars (95% confidence intervals) for each university contained the overall mean (scaled to be zero across all universities). There was no significant variation at the level of the university. (Adapted from Marsh, Martin & Rowe, 2002)

Although PREQ ratings were reliable at the level of individual students, these results are not particularly relevant for the likely application of the PREQ ratings to discriminate between universities. Whereas student evaluation research suggests that PREQ ratings might be reliable at the level of the individual supervisor, the number of graduating PhD students associated with a given supervisor in any one year might be too small to achieve acceptable levels of reliability, and there are important issues of anonymity and confidentiality. There are apparently no comparable studies of the ability of students' evaluations of teaching effectiveness to differentiate between universities or even departments within universities. Nevertheless, these results call into question research or practice that seeks to use students' evaluations of teaching effectiveness as a basis for comparing universities as part of a quality assurance exercise.

2 PROPOSED RESEARCH

This interim report provides preliminary results based on Stage 1 and a three-stage research programme based on the NSS. In this overarching research programme we proposed to use seed funding from Higher Education Academy (herein referred to as “The Academy” and HEFCE to undertake preliminary research presented here and to evaluate the feasibility for a large-scale collaborative study (the present authors, the Academy and HEFCE) study to be funded by ESRC. Hence the plans for subsequent research are dependent upon results from the preliminary analyses and success in obtaining funding from ESRC.

In stage one, we pursued analyses along the lines of those that were conducted with the PREQ on behalf of the Australian Council of Deans and Directors of Graduate Studies and were summarized by Marsh, Rowe and Martin (2002). In preliminary analyses reported here we used data from 2005 and 2006 NSS we conducted analyses to evaluate:

- The factor structure implicit in the design of the NSS responses and its replicability using a combination of exploratory factor analysis, confirmatory factor analysis, and higher-order confirmatory factor analysis;
- The proportions of variance in overall student satisfaction, based on multilevel analyses, that are explained at the level of the individual student, the discipline or field of study, and the university;
- The amount of variance in overall student satisfaction, based on multilevel analyses, that can be explained by background demographic variables (e.g., gender, age, part-time/full-time status) and how controlling for these differences influence proportions of variance at the individual student, the discipline-within-university groups, and the university;
- The reliability of responses at each level of analysis, with particular emphasis on the reliability of differences between universities and fields of study within universities.
- The test-retest reliability and mean stability of the ranking of universities in 2005 with those based on NSS responses in 2006.

The NSS data is uniquely suited to pursue these analyses because of the very large sample size at each level of the proposed analysis. In this stage, we will focus on overall satisfaction as the main outcome variable of interest.

In stage two, we will expand analyses in Stage 1 to focus more specifically on the multidimensionality of the NSS responses and evaluate the usefulness of the different components. We aim to:

- Evaluate the replicability of the NSS factor structure at the levels of the individual student, the discipline-within-university groups, and the university;
- Determine the extent to which student background characteristics are differentially related to different NSS components;
- Evaluate the extent to which the different NSS components are stable over time based on the 2005 and 2006 NSS responses;
- Determine the extent to which the ability of the NSS responses to differentiate between universities and discipline-within-university groups is improved through the evaluation of the multiple NSS dimensions;
- Explore the extent to which different universities and discipline-within-university groups have distinct profiles of different NSS factors (e.g., high in Teaching but low in Resources) rather than being uniformly high, medium, or low across all NSS factors.

In stage three, based on the assumption that NSS responses are psychometrically reasonable, we would like to compare these results based on responses from two consecutive years in relation to potential criterion variables such as:

- Research Assessments (RAEs) of Departments & Universities (these analyses would also provide a preliminary basis for evaluating support for the existence of a teaching-research nexus at the level of field of study and university to counter misinterpretations of the finding that research and teaching effectiveness are nearly uncorrelated at the level of the individual teacher as shown by the Hattie and Marsh, 1996, meta-analysis).
- Rankings from the Quality Assurance Exercises;
- Audits by Professional Organizations (e.g., teaching training assessment by OFSTED for Education);
- Entry standards (as both a potential confounding factor and a potential criterion);
- Staff/student ratios;
- Infrastructure spending on libraries and computers;
- Expenditure on student facilities;
- Percentage of students obtaining honours;

- Graduate destinations (indexed by employment is relevant areas of work);
- Completion rates;
- Some “case studies” to elaborate on how students respond to the survey and what the results mean (as a follow-up to the statistical analyses).

The criterion variables listed here are indicative of the type of variables that are likely to be available. As part of the research this list will be expanded on advice of collaboration with HESA, HEFCE, the Academy, and other organizations with which we will develop collaborative working relations.

3 NSS: BACKGROUND

As part of a revised quality assurance framework for higher education, the first full national student survey in England, Wales and Northern Ireland was conducted in 2005. The National Student Survey forms part of the revised Quality Assurance Framework (QAF) for higher education. Two main aims of the survey were to gather feedback on the quality of students' courses in order to contribute to public accountability and to help inform the choices of future applicants to higher education. The QAF review group has recommended that the survey be run annually for the foreseeable future in order to build a fuller picture and gather data more quickly (HEFCE, 2007).

3.1 Factor structure

Surridge (2006) assumed the six-factor structure underlying the NSS based in part on a preliminary exploratory factor analysis conducted by Richardson (2005), but there appears not to have been confirmatory factor analysis using current approaches to factor analysis. Also, even the preliminary factor analysis was conducted at the level of the individual student, whereas the relevant unit of analysis should be either the institution or the discipline-within-university groups. Thus, for example, although the various NSS scales are reasonably distinct at the level of the individual student, a more relevant question in terms of the actual usage is whether the scales are reasonably distinct at the level of the institution or discipline-within-university groups. Typically, relations between multidimensional scales are substantially higher at higher levels of analysis (i.e., at the institutional level compared to the individual student level).

This has important practical implications, for example, as to whether it is appropriate to interpret a profile of scores (representing the different NSS factors) or a single score.

3.2 Reliability

Issues of reliability of the responses have not been pursued in any report thus far. Indeed, the only brief mention of reliability was in reference to the coefficient alpha estimates of reliability previously reported by the preliminary analyses conducted by Richardson (2005). Several of the scales had marginal coefficient alpha estimates of reliability – less than .80. More importantly, coefficient alpha estimates based on agreement between multiple items are not an appropriate measure of reliability for a multilevel construct in which the relevant unit of analysis is either the institution or discipline-within-university groups. Because the NSS responses are used to compare institutions and discipline-within-university groups, the appropriate measure of reliability should be based upon the extent of agreement among different students within the same institution or the discipline-within-university group. Based on these reliability estimates, error bars – the range of probable error -- around the mean of each university and disciplines within each university can be constructed to facilitate comparisons.

3.3 Ability to Differentiate Between Institutions and Discipline-within-university Groups

Since the NSS responses are used to compare universities and discipline-within-university groups, a critical issue is the extent to which the NSS responses are able to differentiate between these groups. The appropriate analysis is a three level multilevel analysis (level 1 = students; level 2 = discipline; level 3 = institution). Although some aspects of analyses in the Surridge (2006) reports are relevant to this question, it was not directly addressed. Thus, for example, the amount of variance that could be explained by differences between institutions (e.g., only about 2.5% of the variance in NSS responses could be explained by differences between institutions) was considered. A critical issue is how much of these institutional differences could be explained in terms of disciplines represented at each university. It has also been shown by Surridge (2006) that there were systematic differences between disciplines. What was not pursued, however, was whether discipline differences generalized over across institutions or whether the same discipline received substantially higher ratings at some institutions and substantially lower ratings at other institutions. Although either result would be interesting, each would have profound differences on the appropriate use of NSS rating.

3.4 The Role of Individual Student, Course, and Institutional Characteristics

Surridge (2006) provided detailed descriptive results showing how, for example, individual student characteristics contributed to NSS ratings. Although important in their own right and clearly relevant to our aims, there are fundamental differences in our approach and that of Surridge's. Whereas the primary focus of the Surridge reports were on the nature of these differences themselves, our primary foci would be on the extent to which these differences contributed to apparent differences between institutions (and discipline differences within institutions). To the extent that there are substantial effects of student background characteristics, a relevant question is whether or not it is appropriate to correct for such differences prior to ranking universities (or disciplines within universities) – a value-added approach.

It is also important to reiterate that analyses that we propose here are intended to be preliminary analyses that will form the basis of a much larger grant to be submitted to ESRC for a collaborative project with ourselves, HEFCE, and the Academy. However, we are confident that even the preliminary analyses would provide a rich source of information about how best to interpret, use, and apply NSS data. The subsequent ESRC grant, would not involve any funding from either HEFCE or the Academy, and would provide a wonderful opportunity develop an ongoing research program around the NSS that should be of great benefit to you in terms of how best to use the NSS.

4 METHODS

4.1 Participants (and background characteristics)

For 2005 and 2006, demographic information for 285,445 and 278,796 students was obtained. Of these, 171,320 (97,356 females; 73,964 males; mean age: 22.0 years) and 157,371 (93,704 females; 63,667 males; mean age: 21.2 years) students submitted responses to the NSS, representing 140 to 144 universities.

Data for individual student characteristics were based on the student record collected by the Higher Education Statistics Agency (HESA) and were not collected as part of the NSS questionnaire. These data were later matched with the collected NSS data via the unique pupil identifier in the National Pupil Database (HESA, 2007). Characteristics pertaining to the student, discipline and institution constitute a sample of the explanatory variables used in this report (see table 1).

4.2 Materials: NSS Survey Items

The NSS assesses undergraduates' satisfaction with UK universities in seven domains: Teaching, Assessment and Feedback, Academic Support, Organization and Management, Learning Resources, Personal Development, and Overall Satisfaction (see Appendix A). A total of 22 items constitute the survey, e.g. (Teaching: "staff are good in explaining things"; Assessment & Feedback: "Assessment arrangements and marking have been fair"; Support: "I have received sufficient advice and support with my studies"; Organisation: "The timetable works efficiently as far as my activities are concerned"; Resources: "The library resources and services are good enough for my needs"; Personal Development: "The course has helped me to present myself with confidence"; Overall Satisfaction: "Overall, I am satisfied with the quality of the course"). Participants respond to each question on a five-point Likert scale ranging from 1 = "Definitely disagree" to 5 = "Definitely agree". In addition to these five response categories, students were also given the option to mark the question as "Not applicable" and to add further qualitative comments. However, for the purposes of this report, only the quantitative aspects of the responses are analysed. With regards to the questionnaire itself, more detailed psychometric data are presented as part of the results later.

For the purposes of this report, only data from students who completed the NSS were considered. However, even here, there is missing data as many students who completed the NSS did not complete all items. Missing data typically formed a small percentage of the population in the responses of the NSS ranging from 0.27% (question 4, "The course is intellectually stimulating") to 11.89% (question 18, "I have been able to access specialised equipment, facilities or rooms when I needed to") for the 2005 data and ranges from 0.21% (question 22, "Overall, I am satisfied with my course") to 9.77% for question 18 for the 2006 data. Because the amount of missing data was not large, we used the EM method (based on the EM algorithm as operationalised in *SPSS, version 15*) to replace missing data for purposes of these preliminary analyses. It is expected that in

stage 3 of the research, more sophisticated analyses by way of multiple imputation, will be conducted on the missing data.

Table 1. Explanatory variables considered

Student Characteristics	Discipline Characteristics	Institution Characteristics
Gender (Male/Female)	Specificity of discipline classification (e.g. 19 disciplines vs. 41 disciplines)	Number of pupils in each institution
Ethnicity (White/Black/Asian/Other/Unknown)		
Major Source of Tuition Fees (No award/Award/Other)	Number of pupils in each discipline	
Age group (18&under/19/20-21/22-30/31-40/41+)		
Mode of study (Full time/Sandwich/Part time/Other study)		
Domicile (UK/Outside UK)		
Method of response (Web/Post/Email/Other)		
Accommodation (University/Parent/Own Home/Unknown or other)		
Disability (None/Dyslexia/Other/Unknown)		
Qualification aims (Postgraduate/First degree/Vocation/Other qualification less than degree standard)		
UCAS score tariff (<200/200-299/300-399/400-499/500+)		

5 RESULTS

5.1 Factor Structure

According to the a priori design of the NSS, there are 21 items designed to measure six specific course evaluation factors and an additional “overall” satisfaction item. Richardson (2005) provided a preliminary exploratory factor analysis that was consistent with this design. However, there seems to be no rigorous evaluation of the “best fit” empirical factor structure in relation to the a priori design based on confirmatory factor analysis. We began by evaluating the appropriateness of this a priori design.

Exploratory Factor Analysis

Historically, a large number of ad hoc “rules of thumb” have been used to determine the most appropriate number of factors. Typically, the number of factors is roughly about 1/3 of the number of items, although this will vary substantially with the design of the instrument. A more defensible approach is to evaluate the percent of variance explained by successive factors as represented by eigenvalues for each factor.

One widely used guideline is the “eigenvalue greater than 1” rule (retain the number of factors with an eigenvalue greater than 1.0; Kaiser, 1960). The eigenvalue is a mathematical summary of the amount of variance explained by each factor. A strict interpretation of the eigenvalue greater than 1 rule (see table 2) would result in retention of only 5 factors, and this result is consistent across 2005 and 2006 data. However, the 6th and 7th factors also had substantial eigenvalues that were close to 1.0 (see table 2).

Another guideline is called the “scree” plot (Cattell, 1966) in which the number of factors is plotted against the size of the eigenvalues. Cattell suggested that an appropriate number of factors to retain can be identified by the on the scree plot where the decrease of eigenvalues appears to level off to the right of the plot (“scree” is the geological term referring to the debris which collects on the lower part of steep slope where the slope levels off). The idea of the scree plot is to retain a sufficient number of factors so that subsequent factors add little to the variance explained and therefore is the recommended number of factors in the examined questionnaire. From figures 2 and 3 it can be seen that the recommended number of factors is 8 from both 2005 and 2006 results where the line levels off. However, for both years, differences between factor solutions based on 6, 7, and 8 factors are not large (see figure 2).

A third test of the appropriate number of factors is based on a chi-square test of statistical significance and is closely related to subsequent developments in confirmatory factor analysis. The rationale here is to retain enough factors so that the factors are able to explain all of the statistically significant covariation among the items that are the basis of the factor analysis. However, the problem with such tests is that they are substantially influenced by sample size such that small differences are statistically significant (leading

to the decision to retain more factors) when sample sizes are large – as is the case here. Because of the very large sample size, the chi-square test suggested that more than 14 factors should be retained. (see table 2)

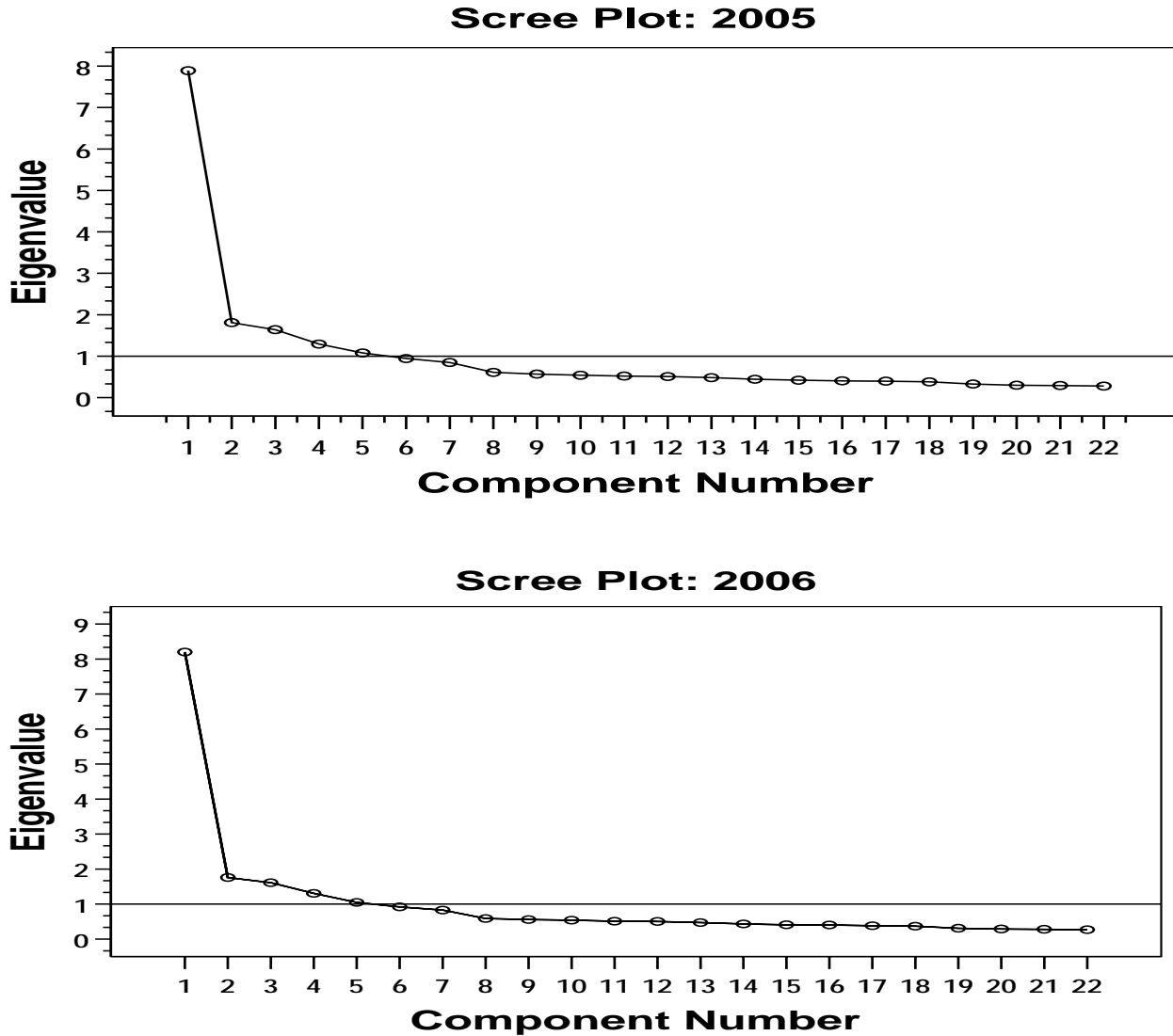


Figure 2. Scree Plot as a basis of determining the number of factors for 2005 and 2006 data

In summary, these guidelines did not provide definitive results for the number of factors. Importantly it needs to be reiterated that all these guidelines are rules of thumb to aid the researcher in the factor structure that provides the most useful interpretation of the data. Critical in this exercise is the substantive evaluation of the various factor solutions in relation to the intended purpose of the study. From this perspective, we evaluated the interpretability of the factor solutions based on different numbers of factors (a principle component extraction followed by an oblique rotation).

When a single factor was retained, all 22 items loaded substantially on this overall factor; factor loadings varied from .3 to .7 except for the overall rating item that had a factor loading greater than .8. This is consistent with the interpretation of the large general overall satisfaction factor (the large eigenvalue for the first factor in table 2 that explains much of the variance in responses to individual items). With the retention of each additional factor, one of the a priori specific factors was identified. When six factors were retained, the six a priori factors were clearly identified (see table 3). However, when seven factors were retained (see table 3), the “Assessment” factor was divided into clearly identifiable components relating to fairness of assessment (i.e., “Criteria used in marking have been made clear in advance; Assessment arrangements and marking have been fair”) and feedback of assessment (“Feedback is prompt; I have received detailed comments on my work; Feedback on my work has helped me clarify things I did not understand”). When we evaluated an 8-factor solution, there were no items that loaded substantially on the eighth factor.

In summary, an evaluation of the various guidelines as to the number of factors and the substantive interpretations of the empirical factor solutions suggest that the “best” solution is either a six-factor solution consistent with the original a priori design or a seven-factor solution in which the original assessment factor is divided into two subcomponents.

Confirmatory Factor Analysis

Confirmatory factor analysis allows the researcher to specify the model to be tested, evaluate the extent to which the posited model fits the data, and to rigorously compare alternative models. For present purposes we focus on three issues: (a) comparison of models positing six specific NSS factors (consistent with the a priori design) and seven factors (consistent with the exploratory factor analysis results); (b) the role of the overall satisfaction (item 22) in terms of how it fits into to NSS factor structure and its relation with specific NSS factors; and (c) the strength of the NSS hierarchy and the extent to which covariation among first-order NSS factors can be explained in terms of a single higher-order satisfaction factor. In evaluating the factor structure for NSS ratings, the overall satisfaction rating (NSS question 22) plays a critical role.

Role of the Overall Satisfaction Rating

Consistent with the student evaluation literature reviewed earlier, it might be appropriate consider the overall rating items as a single “best” indicator of students’ overall satisfaction. Although it might also be appropriate to use a simple average response to all NSS items for this purpose (e.g., the strategy used in the “Times Good University Guide”), this approach implicitly assumes that each of the specific NSS factors (or each NSS item) is equally important. Also, the use of this unweighted average further assumes that there are no additional aspects of satisfaction with educational experience beyond those which have been measured by the NSS. Clearly, each of these implicit assumptions is problematic and requires further consideration.

Table 2. Tests of the variance explained and statistical significance based on different numbers of factors

	<u>Eignevalues</u>		<u>% Variance Explained</u>		<u>Cum% Variance Explained</u>		<u>Chi-Sq ML Tests of Significance</u>		
	<u>2006</u>	<u>2005</u>	<u>2006</u>	<u>2005</u>	<u>2006</u>	<u>2005</u>	<u>DF</u>	<u>2005</u>	<u>2006</u>
1	8.199	7.890	37.267	35.865	37.267	35.865	209	526207.6***	492488.5***
2	1.760	1.811	8.000	8.232	45.267	44.097	188	360723.9***	335925.7***
3	1.612	1.640	7.327	7.453	52.594	51.551	168	240692.8***	226891.2***
4	1.306	1.294	5.938	5.880	58.532	57.431	149	135815.5***	129188.4***
5	1.049	1.080	4.767	4.910	63.300	62.341	131	73524.9***	69184.4***
6	.922	.944	4.190	4.292	67.490	66.633	114	30709.5***	28427.2***
7	.830	.849	3.771	3.857	71.261	70.489	98	10381.3***	9553.4***
8	.587	.611	2.667	2.777	73.928	73.266	83	3242.8***	3149.9***
9	.557	.570	2.532	2.589	76.461	75.855	69	1800.3***	1765.2***
10	.543	.542	2.468	2.462	78.928	78.317	56	897.2***	948.3***
11	.511	.521	2.321	2.367	81.250	80.684	44	491.8***	449.1***
12	.504	.510	2.293	2.317	83.542	83.001	33	309.5***	286.7***
13	.472	.483	2.146	2.195	85.689	85.196	23	212.1***	166.9***
14	.434	.445	1.974	2.021	87.662	87.218	14	120.3***	68.0***

Note. This table gives eigenvalues, variance explained, and cumulative variance explained for factor solution based on varying number of factors retained. The variances extracted by the factors are called the eigenvalues (a name based on the mathematical computations).

Table 3. Factor Pattern Loadings for Six- and Seven-factor Solutions (2005 Data)

	Seven Factor Solution							Six Factor Solution					
	1	2	3	4	5	6	7	1	2	3	4	5	6
1. Staff are good in explaining things.	.678	.004	.027	-.029	.002	-.112	.125	.047	.012	.034	-.020	-.692	-.104
2. Staff have made the subject interesting.	.834	-.005	-.041	.053	-.039	.020	-.030	.013	-.010	-.040	.051	-.845	.022
3. Staff are enthusiastic about what they are teaching	.770	.012	.090	.059	-.003	-.056	-.017	.029	.007	.090	.015	-.782	-.055
4. The course is intellectually stimulating	.713	.020	-.125	-.029	.081	.086	-.012	-.049	.018	-.123	-.075	-.724	.088
5. The criteria used in marking have been made clear in advance	-.066	-.016	-.025	.017	-.011	.027	.908	.652	.039	.002	-.123	.042	.052
6. Assessment arrangements and marking have been fair	.076	.032	.012	.005	.013	-.019	.795	.559	.081	.037	-.129	-.099	.003
7. Feedback on my work has been prompt	-.031	.020	.068	.735	.125	-.024	.075	.749	-.001	.045	-.065	.023	-.054
8. I have received detailed comments on my work	.015	-.015	-.003	.927	-.027	.023	-.023	.860	-.047	-.032	.114	-.020	-.013
9. Feedback on my work has helped me clarify things I did not understand	.034	.015	-.086	.828	-.048	-.028	-.007	.776	-.013	-.111	.123	-.040	-.059
10. I have received sufficient advice and support with my studies	.087	-.016	-.073	.068	-.043	-.708	.065	.102	-.011	-.071	.032	-.092	-.705
11. I have been able to contact staff when I needed to	-.028	.023	.048	-.050	.060	-.871	-.017	-.064	.028	.049	-.069	.026	-.868
12. Good advice was available when I needed to make study choices	-.009	.016	-.075	.054	-.007	-.798	-.015	.034	.017	-.077	.006	.007	-.797
13. The timetable works efficiently as far as my activities are concerned	-.034	.013	-.058	-.018	.817	.059	-.021	-.026	.009	-.072	-.793	.026	.040
14. Any changes in the course or teaching have been communicated effectively	-.021	.005	.037	.051	.813	-.060	.011	.061	.001	.021	-.788	.012	-.080
15. The course is well organised and is running smoothly	.164	-.002	.023	.035	.682	-.092	.064	.079	-.002	.013	-.669	-.177	-.106
16. The library resources and services are good enough for my needs	-.003	.816	-.001	.028	-.029	.044	.010	.033	.816	-.004	.038	.003	.041
17. I have been able to access general IT resources when I needed to	-.005	.865	.021	-.019	-.002	.001	-.004	-.022	.866	.019	.009	.006	-.002
18. able to access specialised equipment, facilities or rooms when I needed to	.007	.797	-.024	-.008	.030	-.055	-.006	-.014	.798	-.027	-.022	-.007	-.059
19. The course has helped me to present myself with confidence	.033	.015	-.840	.014	.002	-.037	.028	.026	.019	-.840	-.009	-.034	-.035
20. My communication skills have improved	-.040	-.001	-.908	.008	-.001	-.001	-.001	.001	.001	-.909	-.004	.040	.000
21. As a result of the course, I feel confident in tackling unfamiliar problems	.034	.011	-.837	.010	.021	-.025	.006	.006	.013	-.837	-.025	-.036	-.024
22. Overall, I am satisfied with the quality of the course.	.398	.040	-.210	.029	.229	-.153	.100	.089	.044	-.210	-.235	-.410	-.154

Extraction Method: Principal Component Analysis. Rotation Method: Oblimin with Kaiser Normalization.

The advantage of the global satisfaction item is that it “finesses” these implicit assumptions by asking students to evaluate directly their overall satisfaction. In this way, each student has the opportunity to subjectively weight all different aspects of his/her educational experience according to the weight given each component by that student; there is no assumption that their global satisfaction rating equally weights all NSS factors. Furthermore, each student is free to consider other aspects of the educational experience that might not be included on the NSS. The relations between this global satisfaction rating and the specific NSS factors also provide one estimate of the relative importance of each factor to students in terms of their overall satisfaction.

The incorporation of the global satisfaction item into the NSS factor structure also provides a complication. Whilst each of the first 21 items designed to measure a specific factor is posited to load only on that one factor (and to have zero loadings on all other factors), this is not reasonable for the overall satisfaction item. Instead, we evaluated a model in which the overall rating item defined a separate factor that was defined by only the single item. Finally, it is also useful to determine how well a single higher-order factor is able to explain covariation among the first-order NSS factors and the extent to which this higher-order satisfaction rating is related to the overall summary rating.

In summary, we began by evaluating a factor structure positing seven factors (the original six specific factors and a seventh overall satisfaction factor; see tables 4 and 5 for 2005 and 2006 models) or eight factors (in which the original Assessment and Feedback factor was divided into two separate factors; see tables 6 and 7).

First-order solution for 2005 and 2006 data

Model 1, based on the a priori (6-factor) design of the NSS, provided an excellent fit to the data in relation to traditional indices of fit (see table 4). Furthermore, the goodness of fit was nearly the same across 2005 and 2006 responses. Inspection of the parameter estimates (See table 5) reveals that all factors are well defined. The factor loadings are for all the specific NSS factors are statistically significant and substantial. Whilst all factors are positively correlated (.28 to .76 across the two solutions), none of the factor correlations approaches 1.0 (which might suggest that the factors should be combined).

For both the 2005 and 2006 data, the highest correlation is between the Overall Satisfaction factor and the Teaching factor (.75 in 2005 and .76 in 2006). Indeed, there is a logical pattern of relations between the overall satisfaction factor and the six specific factors. Whilst all six factors are substantially correlated with the overall rating factor, the correlations are highest with the Teaching factor, very high for the Support and Organisation factors, somewhat lower for the Personal Development and Assessment factors, and lowest for the Resources factor.

Table 4. Summary of goodness of fit for all models

<u>MODEL</u>	<u>χ^2</u>	<u>DF</u>	<u>TLI</u>	<u>RNI</u>	<u>RMSEA (90% CI)</u>	<u>SRMR</u>	<u>Description</u>
Model 1: 22 items, 7 FO factors							
1a 2005	78100	189	.981	.984	.0491 (.0488-.0493)	.035	N=171320
1A 2006	77983	189	.981	.984	.0511 (.0508-.0514)	.037	N=157371
Model 2: 22 items, 8 FO factors							
2a 2005	44223	182	.988	.991	.0376 (.0373-.0379)	.027	N=171320
2b 2006	45353	182	.988	.990	.0397 (.0394-.0400)	.029	N=157371
Model 3: 22 items, 7 FO factors, 1 HO factor							
3a 2005	94048	203	.979	.981	.0519 (.0517-.0522)	.041	N=171320
3b 2006	93870	203	.979	.982	.0541 (.0539-.0544)	.043	N=157371
Model 4: 22 items, 8 FO factors, 1 HO factor							
4a 2005	71796	202	.983	.985	.0455 (.0452-.0458)	.038	N=171320
4b 2006	72378	202	.983	.985	.0476 (.0474-.0479)	.040	N=157371
Model 5: 22 items, 8 FO factors CFA Invariance across 2005/2006							
5a	89577	364	.988	.991	.0386 (.0384-.0388)		No Invariance (INV)
5b	90007	386	.989	.990	.0376 (.0374-.0378)		INV=factor loadings (FL)
5c	90254	414	.989	.990	.0363 (.0361-.0365)		INV= FL, Factor var/covar
5d	91324	435	.990	.990	.0357 (.0355-.0359)		Total invariance

Note. RNI = relative noncentrality index, TLI = Tucker-Lewis index, RMSEA = root mean square error of approximation; 90% CI = 90% confidence interval for the RMSEA, DF = degrees of freedom.

Table 5. Model 1: NSS 7 factors 22 items: 2005, 2006 data

Factor Loadings

	Teach		Assess		Support		Organ		Resource		Develop		Overall	
	2005	2006	2005	2006	2005	2006	2005	2006	2005	2006	2005	2006	2005	2006
Q01	.71	.72												
Q02	.76	.77												
Q03	.68	.69												
Q04	.66	.68												
Q05			.53	.54										
Q06			.57	.58										
Q07			.70	.70										
Q08			.79	.79										
Q09			.79	.80										
Q10					.78	.79								
Q11					.69	.69								
Q12					.76	.77								
Q13							.55	.57						
Q14							.72	.73						
Q15							.87	.87						
Q16									.65	.64				
Q17									.76	.76				
Q18									.77	.75				
Q19											.85	.86		
Q20											.81	.82		
Q21											.82	.83		
Q22													1.0	1.0

Factor Correlations

	Teach		Assess		Support		Organ		Resource		Develop		Overall	
	2005	2006	2005	2006	2005	2006	2005	2006	2005	2006	2005	2006	2005	2006
2005														
2006														
Teach	1.00													
Assess	.62	.62	1.00											
Support	.70	.72	.66	.68	1.00									
Organ	.62	.63	.56	.55	.65	.66	1.00							
Resource	.30	.33	.28	.31	.41	.43	.37	.40	1.00					
Develop	.60	.62	.45	.47	.55	.57	.44	.44	.34	.36	1.00			
Overall	.75	.76	.57	.57	.68	.69	.68	.69	.35	.37	.60	.61	1.00	1.00

Higher-Order Factor Loadings

HO Factor	Teach		Assessment		Support		Organ		Resource		Develop		Overall	
	2005	2006	2005	2006	2005	2006	2005	2006	2005	2006	2005	2006	2005	2006
HO Factor	.86	.87	.71	.71	.83	.84	.77	.77	.43	.45	.67	.69	.86	.86

Note. Results are based on an a priori 7-factor model applied separately to data from 2005 and 2006. Parameter estimates are presented in completely standardised format.

However, a more detailed evaluation of the Assessment factor revealed some potential problems. Whereas the all five factor loadings are substantial, items 5 and 6 (dealing with fairness) have much smaller factor loadings than the remaining three items. Furthermore, an inspection of modification indices¹ indicated that correlations among the first two assessment items and correlations among the last three assessment items were substantially higher than could be explained in terms of the single Assessment and Feedback factor. This pattern of results is consistent with earlier results suggesting the need to separate the Assessment and Feedback factor into two separate factors.

Model 2 posits 7 specific NSS factors (with the assessment factor divided into two separate factors) and an eight factor based on the overall rating item. Model 2 also provided an excellent fit to the data (table 4). For both the 2005 and 2006 data, the fit for Model 2 was significantly better (based on the chi-square goodness of fit tests). Even for the indices (TLI and RMSEA) that penalise lack of parsimony (the 8-factor model is less parsimonious than the 7-factor model), the goodness of fit for the 8-factor model is better than the corresponding 7-factor model and this result is consistent across 2005 and 2006. Furthermore, a more detailed inspection of the parameter estimates supports the superiority of the 8-factor solution (see table 6). In particular, the factor loadings associated with each of the five assessment items is systematically higher in the 8-factor solution than in the corresponding seven-factor solution. Although the two assessment and feedback factors are substantially correlated (.67 in 2005, .68 in 2007), these correlations are significantly (and meaningfully) less than 1.0 (the correlation based on the assumption that the two factors reflect the same underlying construct).

In summary, although there is good support for both the 7- and 8-factor solutions, support for the 8-factor solution is clearly stronger based upon the goodness of fit of the factor solutions examined here.

¹ Modification indices are another source of identifying model mis-specification. It represents the decrease in the value of the chi-square when the parameter is estimated in a revised model. Thus, it is useful to estimate the parameters associated with the largest modification indices.

Table 6. Model 2: NSS 8 factors 22 items: 2005, 2006 data

Factor Loadings																
	Teach		Fair		Fdbck		Support		Organ		Resource		Develop		Overall	
	2005	2006	2005	2006	2005	2006	2005	2006	2005	2006	2005	2006	2005	2006	2005	2006
Q01	.71	.72														
Q02	.76	.77														
Q03	.68	.69														
Q04	.66	.68														
Q05			.66	.67												
Q06			.75	.76												
Q07					.69	.68										
Q08					.82	.82										
Q09					.82	.83										
Q10							.78	.80								
Q11							.69	.69								
Q12							.76	.71								
Q13									.55	.57						
Q14									.72	.73						
Q15									.87	.87						
Q16											.65	.64				
Q17											.76	.76				
Q18											.77	.75				
Q19													.85	.86		
Q20													.81	.82		
Q21													.82	.83		
Q22															1.0	1.0
Factor Correlations																
	Teach		Fairness		Feedback		Support		Organ		Resource		Develop		Overall	
	2005	2006	2005	2006	2005	2006	2005	2006	2005	2006	2005	2006	2005	2006	2005	2006
Teach	1.00															
Fair	.57	.59	1.00													
Feedback	.58	.58	.67	.68	1.00											
Support	.70	.72	.62	.63	.62	.63	1.00									
Organ	.62	.63	.56	.57	.50	.50	.65	.66	1.00							
Resource	.30	.33	.32	.34	.25	.27	.41	.43	.37	.40	1.00					
Develop	.60	.62	.40	.43	.42	.44	.55	.57	.44	.44	.34	.36	1.00			
Overall	.75	.76	.56	.57	.52	.53	.68	.69	.68	.69	.35	.37	.60	.61	1.00	
Higher-Order Factor Loadings																
HO Factor	Teach		Fairness		Feedback		Support		Organ		Resource		Develop		Overall	
	2005	2006	2005	2006	2005	2006	2005	2006	2005	2006	2005	2006	2005	2006	2005	2006
HO Factor	.85	.86	.70	.72	.67	.67	.83	.84	.77	.77	.43	.45	.67	.68	.85	.85

Note. Results are based on an a priori 7-factor model applied separately to data from 2005 (Model 2a) and 2006 (Model 2b). Parameter estimates are presented in completely standardised format. Also shown are the higher-order factor loadings relating each first-order factor to the global satisfaction factor (Models 4a, 2005; Model 4b, 2006).

Higher-order solution for 2005 and 2006 data

In both Models 1 and 2, there were substantial correlations among the specific satisfaction factors. This suggests the possibility of a higher-order, global satisfaction factor. In order to test this possibility, we evaluated a higher-order factor in which all the correlations among the seven first-order factors in Model 1 and the 8 first-order factors in Model 2 could be explained in terms of a single higher-order factor (see Models 3 and 4 in table 4; also see figure 3).

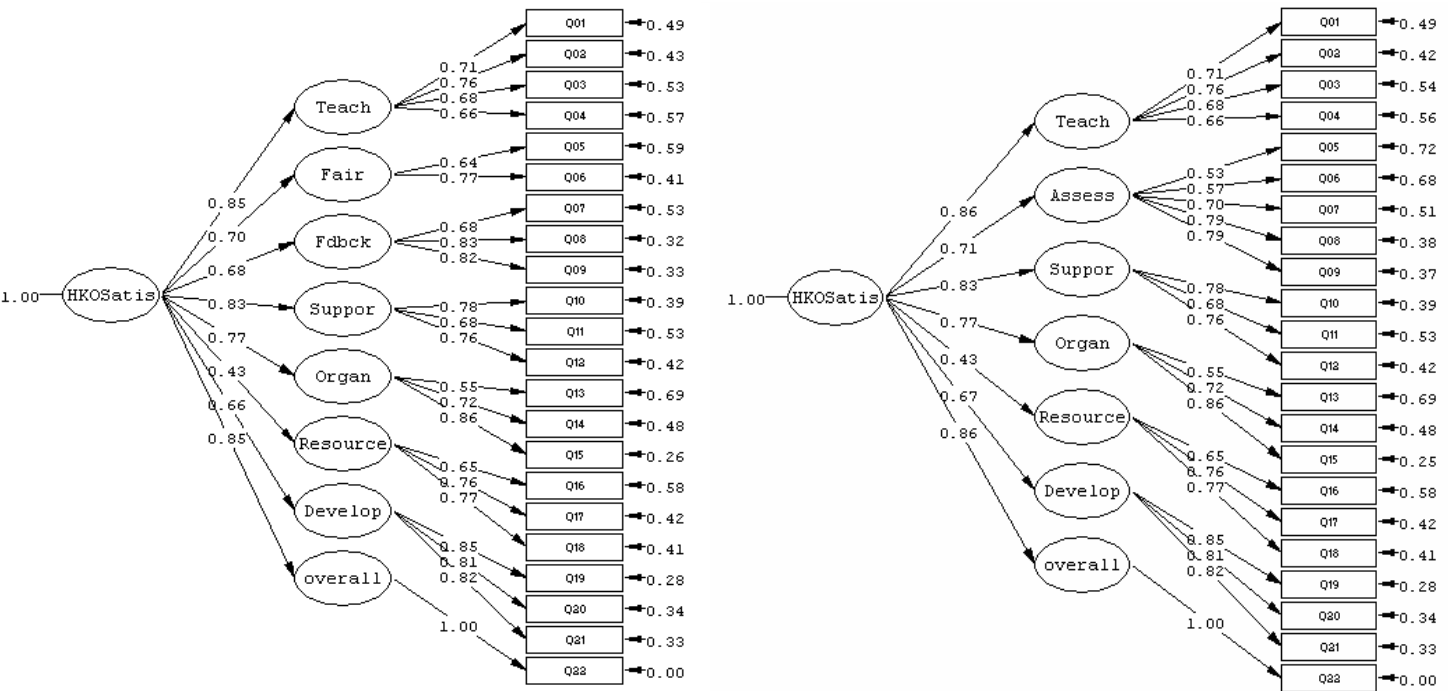


Figure 3. 7-factor and 8-factor structures with 1 higher-order factor of overall satisfaction.

In the evaluation of the higher-order factor solutions, we begin with the goodness of fit statistics. Both the higher-order models (Models 3 and 4 in table 4) resulted in very good fit indices. Again, however, the fit of the 8-factor model (seven specific and one overall satisfaction factor) was better than the 7-factor model. Although the higher-order models provided a very good fit to the data, in each case the fit of the higher-order model was worse than the corresponding first-order model. This change in goodness of fit was highly significant in a statistical sense (based on the change in chi-square values), but more modest in terms of the goodness of fit indices. Furthermore the pattern of results was highly similar for the 2005 and 2006 data.

The intent of the higher-order factor structure is to explain covariation among the first-order factors. Results based on Models 3 and 4 suggest that the single higher order factor is

reasonably successful in accomplishing this task for both models based on 7 and on 8 first-order factors. Again, these results are consistent across the 2005 and 2006 data sets.

An alternative perspective is how successful the higher-order model is at explaining variation in the first order factors. The amount of variance in the first order factors explained by the higher-order factor is based on the second-order factor loadings (see “Higher-Order Factor Loadings” in tables 5 and 6). In each of the four HO models (7 and 8 FO factors, for 2005 and 2006 data) there is a logical and highly consistent pattern of relations between the HO global satisfaction factor and the FO. Whereas all HO factor loadings relating the HO factor to each of the FO factors, the loadings are highest for the overall satisfaction factor and the teaching factor, nearly as high for the supportiveness factor, very high for the organisation factor, somewhat lower for the personal development and assessment factors, and lowest for the resources factor. Whereas many of these HO factor loadings are substantial, even a HO factor loading of .71 means that approximately 50% ($.71 \times .71 = .5041$) of the variance in the FO factor is explained by the HO factor. From this perspective, there is much reliable variance in the FO factors that is unexplained by the HO factor – particularly for the resource, development, assessment and feedback factors where less than half of the variance can be explained in terms of the HO satisfaction factor.

In summary, results based on the HO factor analysis suggest that there is a strong global satisfaction factor associated with the first-order NSS factors. Whereas this single factor is able to explain much of the covariation among the FO factors, there are substantial amounts of variance in each of the FO factors that cannot be explained by the HO factor. These results have important practical implications for how the NSS results are summarised. The success of the HO factor suggests that it is reasonable to summarise the NSS results in terms of a single score – an appropriately weighted average of the first-order factors or responses to the overall satisfaction rating. However, in doing so, considerable information in the first-order factors is lost. A critical issue not addressed by these results is the importance of information in the FO factors that would be lost by using only a single overall satisfaction score. We will address this issue in part in subsequent analyses in which we evaluate the extent to which individual student, departmental, and university characteristics are differentially related to the various NSS first-order factors. Although beyond the scope of this study, the answer to this question clearly varies with the use that is made of the NSS results. For summative purposes, results based on a single global satisfaction factor might suffice, but for formative feedback to universities and their departments, a profile of specific factors that illustrates relative strengths and weaknesses in different areas is likely to be more useful. Further research is needed to evaluate the usefulness and appropriateness of a profile of specific NSS factors for use by students in selection of universities – rather than a single overall summary item currently used in publications such as the Times Good University Guide.

Invariance of factor solutions across the 2005 and 2006 data

A critical issue is the extent to which the underlying factor structure is consistent across the 2005 and 2006 data. This issue has important practical implications in terms of the usefulness of the ratings. Implicit in the collection of data in different years is the assumption that the profile of scores is comparable from one year to the next – to determine the extent to which universities and their departments have improved over time. A critical assumption underlying this use of the ratings is the invariance of the factor structure over time – the extent to which the same

underlying factors are measured in each year that data are collected. Results presented thus far suggest that there is a high degree of similarity in the 2005 and 2006 data sets. In this section we pursue tests of factorial invariance to test this suggestion more formally.

Given appropriate data from two successive years, it is important to evaluate the extent to which the responses to the 22 NSS items are associated with similar factors (the same factors exist), relations between items and latent factors are similar (invariance of the factor loadings), relations among the different factors are similar (invariance of the factor correlations and factor variances), and measurement error is similar (invariance of the uniquenesses). In pursuing this issue, we conducted a series of multiple-group CFAs comparing the factor structures based on responses with the same 22 NSS items by students who completed the instrument in 2005 and 2006. Of particular interest are the factor loadings that constitute the minimum condition of factorial invariance but also the factor variance– covariance matrix.

In the initial Model 5a (see Table 4), the a priori factor structure is fitted separately for each group separately and no invariance constraints are imposed. The fit of this “no invariance” model is excellent according to each of the fit indices (e.g., RMSEA = .0386). In Model 2, the factor loadings are constrained to be equal across the two groups. This is broadly taken to be the minimum requirement of factorial invariance. The RMSEA is actually somewhat better than that for Model 5a, providing strong support for invariance of factor over 2005 and 2006. In Models 5c and 5d, additional constraints are imposed on factor variances, factor covariances, and measurement uniquenesses. In each case the fit of the more constrained model is marginally better according to fit indices that take into account parsimony (the TLI and RMSEA). Indeed, the results are remarkable in the strength of the invariance. Typically, the imposition of invariance constraints results in a poorer goodness of fit – even for fit indices that control for parsimony – so that the critical question is how much of the good fit of the no invariance model is lost when invariance constraints are added. Here, the results indicate the fit actually improves with the imposition of invariance constraints. In summary, the results provide remarkably strong support for the invariance of the factor NSS factor structure over time.

5.2 Multi-level Analyses: Appropriate Unit of Analysis

For purposes of the multilevel analyses presented in this section, we focus on students' overall satisfaction with their university experience. This focus is appropriate in that some measure of overall satisfaction is the basis of most benchmarking exercises based on NSS responses. Also, overall satisfaction is broadly appropriate to all settings whereas some of the specific factors are not.

The multilevel modelling perspective that is the focus of this research is important for higher education research (see Ethington, 1996; Marsh, Rowe & Martin, 2002). Almost all data in higher education are inherently multilevel, although—at least historically—this feature of the data has been typically ignored. Depending on the application, the different levels of analysis might include countries, geographic regions or states, different universities, faculties or departments within universities, and individual students or academic staff. As illustrated by Marsh et al. in their Australian research with the PREQ and in the present investigation, research, policy questions, data, and statistical analyses that are appropriate at one level of analysis may be inappropriate or even misleading when evaluated at another level of analysis.

What is the appropriate unit of analysis for evaluating NSS responses? In student evaluation research discussed earlier it is well established that the appropriate unit of analysis is the individual teacher teaching a particular class rather than the individual student. From this perspective, the most appropriate estimate of reliability for class-average responses is defined by the extent of agreement among students within the same class and not the reliability of responses by individual students. Thus, reliable items are ones for which there is substantial agreement among students within the same class and substantial differences in the mean ratings for different classes. Relations with validity criteria and potential biases based on responses by individual students – instead of class-average responses – are largely irrelevant and not given much attention in this research literature.

For the NSS data it is clear that the individual student is not the appropriate unit of analysis. However, given the focus of the ratings on overall educational experience, it is neither feasible nor, perhaps, even appropriate to consider the individual teacher as a unit of analysis. In this respect, evaluation of NSS responses is very different for studies based on university students' evaluations of teaching effectiveness. A particularly relevant basis for determining the appropriate unit of analysis is the target unit at which the ratings are intended to be used. From this perspective the intended unit of analysis appears to be the entire university, or specific discipline-within-university groups.

As emphasised by Surridge (2006; also see earlier discussion of Marsh, Rowe & Martin, 2002), “the structure of the NSS data is innately hierarchical – that is students do not take courses in a vacuum, they take them at specific institutions. Moreover, the characteristics of the course and the institution of study are expected to have an impact on the responses to the survey”. Thus, whenever the data has a multilevel structure, multilevel modelling should always be the statistical technique of choice. It is generally inappropriate to pool responses of individuals without regard to groups as in single-level analyses, unless it can be shown that there are no group differences or effects. In the present investigation, this would mean that single-level analyses are inappropriate

unless it can be shown that there are no systematic differences between universities and discipline-within-university groups in relation to NSS responses – in which case the NSS ratings would not be appropriate for benchmarking comparisons.

Rationale and Results For Multilevel Analyses In the Present Investigation

In the present investigation, we focus on a three-level model in which individual student is the lowest level, university is the highest level, and discipline-within-university groups is the intermediate level and the dependent variable is the Overall Satisfaction factor. In all cases each of these effects is appropriately considered to be a random effect (see Marsh, Rowe & Martin, 2002; also see Bryk & Raudenbush, 1992; Goldstein, 1995).

The classification of students and universities as random effects is straight-forward, but the classification of academic discipline is more complicated. If the concern is to test for the statistical significance of differences between a fixed set of disciplines, then it might be appropriate to consider these as fixed effects (as well as random effects associated with discipline). Thus, for example, if the focus of the NSS was to compare satisfaction of students in Psychology (averaged across all universities) with those in economics (also averaged across all universities), then it would be appropriate to consider discipline as a fixed effects (in which two of the different discipline categories were, for example, Psychology and Economics). Although this is a potentially interesting and relevant question, it is not the focus of the present investigation and apparently not the main focus of the NSS. Rather, the benchmarking focus of the NSS is to compare, for example, Psychology courses across different universities and to compare Economics courses across universities – not necessarily to compare Psychology courses with Economics courses. Hence, the more relevant question is the extent to which satisfaction ratings for a Psychology course at a particular university differs systematically from those of Psychology courses at different universities rather than the average satisfaction for Psychology courses averaged across all universities.

More generally, the focus of the present investigation is on the extent to which there are substantial amounts of variance due to discipline-within-university groups, treating discipline as a random effect. Marsh, Rowe and Martin (2002) argued that these two perspectives – discipline as a random effect and as a fixed effect -- are not mutually exclusive. More specifically, they argued that it is useful to consider discipline as both a fixed and a random effect in the same model – an approach that we also use in the present investigation. We also note that the issue of discipline in the present investigation is further complicated by the fact that discipline can be represented by alternative classification schemes that differ in terms of the discipline categories (e.g., 19, 41, or 150 discipline groups for 2005; 19, 41 or 565 discipline groups for 2006).

Variance components

In Model 1 (table 7) we began by considering an overly simplistic two-level model (level 1 = students, level 2 = university) in which discipline is ignored. Model 1 is a variance component model (sometimes referred to as a “null model”) in that there are no fixed effects – only random effects due to students and university. For both 2005 and 2006 NSS responses, slightly more than 4% of the variance in NSS responses can be explained in terms of university. The remaining

(residual variance) is due to differences among student within each university – lack of agreement among students from the same university in terms of their overall satisfaction with their educational experience. Particularly given the very large sample sizes, the university effect – even though small in absolute terms – is highly significant from a statistical perspective. In this sense, the NSS ratings can be said to differentiate between universities.

Model 2 (table 7) is more appropriate in that it is based on a three level model (level 1 = students, level 2 = discipline-within-university groups, level 3 = university) consistent with hierarchical structure of the NSS data. Although the results vary depending on the number of discipline categories that are considered, there is a consistent pattern in the results across 2005 and 2006 data. In all cases, the amount of variance associated with universities drops (from about 4% to about 3%). Hence, variance components based on the two-level model were systematically biased in the direction of exaggerating differences between universities that were more appropriately explained in terms of differences between discipline-within-university groups. Not surprisingly, the variance components associated with discipline increase systematically with increases in the number of discipline categories. In the 2005 data, for example, the amount of variance explained by discipline-within-university groups is 7.1% (150 discipline categories), 6.0% (41 categories) and 4.9% (19 categories). Although this suggests that more discipline categories should be considered, there is an important limitation to this strategy. In particular, many discipline groups within a particular university may be so small that the results are not adequately reliable to be considered in isolation. Clearly there is a trade-off between number of students within each discipline-within-university groups and the number of discipline classifications that are considered. We address this issue in more detail latter in the study (see section entitled “Differentiation as a function of Error and Reliability”).

In Model 3 (table 7) we consider a three level model (like Model 2) that has discipline-within-university groups as a random effect, but also includes discipline as a fixed effect. For present purposes, our focus is not on discipline differences (averaged across universities) per se. Nevertheless, because these discipline differences are interesting in their own right, the results of these fixed effects are summarized in Appendix 4 (also see SurrIDGE, 2006, 2007 for a detailed description of these discipline differences and their implications). Rather, the focus of the present investigation is on how controlling for these fixed effects of discipline change the variance components due to the random effects of university and discipline-within-university groups. This is appropriate if the focus is on differences between universities (controlling for the fact that universities differ in terms of disciplines offered) and the comparison disciplines within one university with those at other universities. Not surprisingly, controlling for discipline fixed effects further reduced the variance components associated with university but had even larger effects on variance components associated with discipline-within-university groups.

In Model 4 (table 7) we control for a wide range of individual student characteristics (e.g., gender, ethnicity, age, disability, prior attainment). Again, our focus

Table 7. Residual variance associated with university, discipline and student level of analysis when considering differences between disciplines within and between universities (standard errors in parentheses).

Year	Discipline Classification	Level of Analysis	Model 1	Model 2	Model 3	Model 4	Model 5
2005	(A) 19	University	0.041(0005)	0.032(0.005)	0.027(0.004)	0.034(0.005)	0.029(0.004)
		Discipline		0.049(0.002)	0.033(0.002)	0.048(0.002)	0.033(0.002)
		Student	0.945(0.003)	0.922(0.003)	0.923(0.003)	0.914(0.003)	0.914(0.003)
	(B) 41	University		0.030(0.004)	0.021(0.003)	0.032(0.005)	0.023(0.003)
		Discipline		0.060(0.002)	0.043(0.002)	0.059(0.002)	0.043(0.002)
		Student		0.898(0.003)	0.899(0.003)	0.890(0.003)	0.890(0.003)
	(C) 150	University		0.031(0.005)	0.020(0.003)	0.034(0.005)	0.022(0.003)
		Discipline		0.071(0.002)	0.050(0.002)	0.070(0.002)	0.050(0.002)
		Student		0.902(0.003)	0.902(0.003)	0.893(0.003)	0.893(0.003)
2006	(A) 19	University	0.043(0.005)	0.031(0.005)	0.026(0.004)	0.034(0.005)	0.029(0.004)
		Discipline		0.050(0.003)	0.036(0.002)	0.051(0.003)	0.037(0.002)
		Student	0.959(0.003)	0.923(0.003)	0.923(0.003)	0.914(0.003)	0.914(0.003)
	(B) 41	University		0.032(0.005)	0.023(0.003)	0.035(0.005)	0.026(0.004)
		Discipline		0.067(0.003)	0.050(0.002)	0.068(0.003)	0.051(0.002)
		Student		0.906(0.003)	0.906(0.003)	0.896(0.003)	0.897(0.003)
	(C) 565 ^a	University		0.035(0.005)	0.023(0.003)	0.039(0.005)	0.025(0.004)
		Discipline		0.092(0.003)	0.075(0.002)	0.094(0.003)	0.076(0.002)
		Student		0.887(0.003)	0.888(0.003)	0.878(0.003)	0.878(0.003)

Description of Models: (1) 2-level (university, student) model with no fixed effects; (2) 3-level (university, discipline, student) model with no fixed effects; (3) 3-level controlling for discipline fixed effects; (4) 3-level controlling for student characteristics fixed effects; (5) 3-level controlling for discipline and student characteristics fixed effects. Each model was applied separately to data from each year (2005, 2006) based on different discipline classifications.

^a41 categories of discipline was added as a fixed effect and not 565 categories.

is not on the effects of these individual student characteristics, but the results of these fixed effects are summarized in Appendix 5 (also see SurrIDGE, 2006, 2007 for a detailed description of these effects of individual student characteristics and their implications). More specifically, our focus is how controlling for these individual student characteristics influence variance components associated with the university and discipline-within-university groups rather than the nature of these student effects per se (see Appendix 5 for a summary of these student effects). Consistent with this perspective, SurrIDGE (2006) concluded that it is important to explore the need to control for student characteristics when using NSS ratings to compare universities or discipline-within-university groups in order to avoid potential biases. Many of these individual student characteristics are highly significant from a statistical perspective (due in part to the very large Ns). However, controlling for these characteristics has only very small effects on the variance components. However, these differences are small – reflecting that fact that overall satisfaction ratings are not substantially influenced by these individual student characteristics.

Finally, in Model 5 (table 7) we simultaneously control for fixed effects associated with discipline (as in Model 3) and student characteristics (as in Model 4). However, because student characteristics have such small effects, these variance components in Model 5 are similar to those in Model 3.

In summary, we have explored the ability of the NSS responses to the overall satisfaction rating to differentiate between universities and between discipline-within-university groups. At the level of the university, the overly simplistic Model 1 provides the most optimistic picture, indicating that differences due to university explain slightly more than 4% of the variance for both 2005 and 2006 responses. However, when discipline-within-university groups are included as a random effect in Model 2, this variance component drops to about 2.5%, depending on the discipline classification that is considered. Interestingly, the inclusion of a wide variety of student characteristics – either by themselves (Model 4) or in combination with discipline fixed effects (Model 5) – has little effect on university variance components. Thus, differences between universities explain about 2.5% of the variance in the NSS overall satisfaction rating.

It is important to re-emphasise that our analyses were based on the overall satisfaction rating. Thus, institutional differences associated with specific NSS components (e.g., Learning Resources and Infrastructure) might be larger or smaller than those based on overall satisfaction. It might also be argued that specific NSS components – or even a profile of the components – would be useful for benchmarking universities. Hence, extension of these results to include specific NSS components as well as the overall satisfaction rating, is a potentially important direction for further research.

Differences due to discipline-within-university groups explain substantially more variance than differences due to university. However, the amount of variance explained by discipline-within-university groups varies substantially with the discipline classification that is considered, ranging from about 5% for the least detailed breakdown (Model 2 with 19 discipline categories) to over 9% for the most detailed classification in

the 2006 data. Some of this differentiation is due to differences in discipline that generalize across universities (as reflected in Model 3 that included fixed effects associated with discipline); variance components in Model 3 decreased between 19% and 30%, depending on the discipline classification. However, even after controlling for discipline fixed effects, much of the differences associated with discipline-within-university groups remained. Again, controlling for fixed effects associated with individual student characteristics had little effect on variance components for random discipline effects. Thus, differences between discipline-within-university groups explain between 3.5% and 7.5% of the variance in the NSS overall satisfaction rating, depending on number of discipline categories are considered. Whilst these results might suggest that more detailed discipline classifications should be used, there is a complex balance between the extent of differentiation between groups and the probable error in ratings for each group that we now address.

5.3 Differentiation as a Function of Error and Reliability

The variance components in Table 7 provide a summary of how well the NSS ratings are able to differentiate between universities and between disciplines-within-university groups—the main focus of a benchmarking exercise. The larger the variance components are, the better the response are able to differentiate between groups and the more useful they are for benchmarking purposes. Here we explore alternative perspectives on this same issue.

Graphic Representation of Differentiation: Caterpillar Plots

The “caterpillar” plots give another perspective on the ability of the NSS ratings to differentiate between universities (see figures in Appendices 2 & 3). Based on Models 2 and 5 (table 7), we constructed corresponding caterpillar plots to reflect differentiation among universities (Appendix 2A-L) and differentiation among discipline-within-university groups (Appendix 3A-L). We consider the initial caterpillar plot in detail and then briefly summarise key aspects of the remaining plots.

The caterpillar plot 2a (see Appendix 2A) illustrates differentiation among universities in 2005 based on the three-level model (L3=universities, L2=discipline based on 19 classifications, L1=students) with no fixed effects. For purposes of this plot all universities are ordered (from lowest to highest) in terms of their mean response to the NSS overall satisfaction rating. For each university, the mean rating and an error bar (95% confidence interval around the mean) are presented. Because these are residuals based on standardised ($Mn = 0$, $SD = 1$) values, the differences are in terms of SD units with a mean of zero across all universities. Universities at the extreme left side of the plot have error bars that are completely below the grand mean (zero); they are significantly below average. Those on the right end of the plot are significantly above average. The majority of the universities in the middle have error bars that overlap with zero and thus do not differ significantly from the grand mean. Here, increased differentiation reflects larger differences between universities and smaller error bars.

The length of the error bars is important in providing a range of probable error in assessing satisfaction levels for each university. This probable error reflects a combination of the extent agreement among students within a given university and the number of students responding from that university. If there is no agreement among students within a particular university or the number of students responding from that university is small, then the range of probable error is likely to be large and interpretations should be made cautiously – if at all. The average number of students responding within each of the 141 universities is over 1,000, and values vary substantially from close to 5000 for the Open University to less than 100 students responding for a small group of universities (see figure 4).

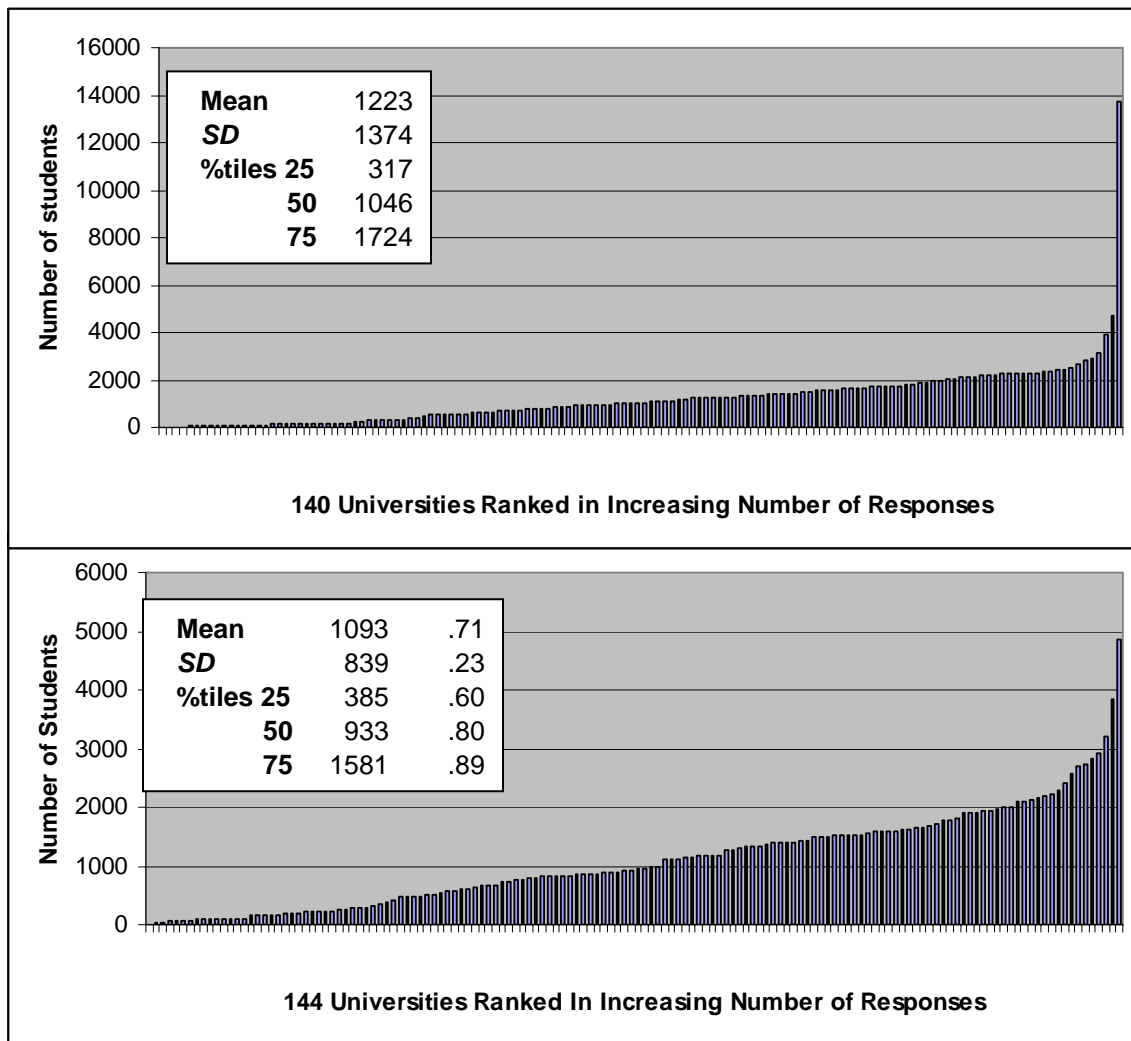


Figure 4. Distribution of sample sizes across universities for 2005 (above) and 2006 (below) data

To illustrate this issue of probable error and the sample sizes, we have highlighted the universities with the smallest number of responding students. In each case, these

universities have large error bars (the six error bars highlighted in red in Appendix 2A). The largest error bar is for the “University of London (Institute & Activities)” based on responses by only 11 students, followed by the “Institute of Education,” “Courtauld Institute of Art,” and the “Royal Academy of Music” and the “Royal College of Music” and the “Royal Northern College of Music” – each with between 25 and 58 respondents.

For caterpillar plots of the differentiation among universities (Appendix 2A-L) the plots are reasonably similar. The means are somewhat less differentiated in Model 5 than Model 2, but the error bars also tend to be somewhat smaller for Model 5 than Model 2. Because these two features tend to cancel out each other, the differentiation between universities does not vary substantially in Models 2 and 5 (as is also evident in the variance components for models 2 and 5 in table 7).

The caterpillar plots for discipline-within-university groups (Appendix 3A-L) differ from those based on universities in a number of important ways. In particular, there are obviously many more discipline-within-university groups and the actual number varies substantially with the discipline classification. Thus, for example, for the 150-discipline classification (2005 data) there are more than 4000 discipline-within-university groups. [If every university had all 150 disciplines, the number of discipline-within-university combinations would be $21,150 = 141 \text{ universities} \times 150 \text{ disciplines}$, but many universities have only a small number of this possible disciplines].

The means associated with different disciplines-within-university groups (Appendix 3A-L) are substantially more extreme than the corresponding university means (Appendix 2A-L). For example, university means were mostly between $\pm 0.4SD$ from the mean (Appendix 2A), whereas the corresponding range of means for discipline-within-university groups is $\pm 1.0SD$ (Appendix 3A). However, the error bars for the means based on discipline groupings are also much wider. This is due in large part to the substantial influence of the number of students on probable error. The error bars grow systematically larger as the number of groups increases (and the average number of students within each group necessarily decreases). For this reason, there are only a relatively few discipline-within-university groups that are significantly different from the grand mean across all students (i.e., have an error bar that does not include the grand mean across all groups). The variance component associated with discipline increases when there are more disciplines (see Model 2, table 7) as is also evident from the more extreme means in the caterpillar plots (e.g., Appendix 3a with 19 categories vs. Appendix 3C with 150 categories). However, because the sizes of the error bars increases substantially as the number of discipline-within-university groups increase, there is a smaller proportion of groupings that differ from the mean when the number of groupings is larger.

Multilevel Reliability

Previous consideration of the reliability of NSS responses has been limited primarily to coefficient alpha estimates of reliability for scale scores (Richardson, 2005). Several of the scales had marginal coefficient alpha estimates of reliability – less than .80. More importantly for purposes of the present investigation, coefficient alpha estimates based on agreement between multiple items are not an appropriate measure of reliability for a multilevel construct in which the relevant unit of analysis is either the university or the discipline within the university. Because the NSS responses are used to compare universities and discipline-within-university groups, the appropriate measure of reliability should be based upon the extent of agreement among different students within the same group (university or discipline-within-university group) and the number of students in the group.

In multilevel analysis, the extent to which groups differ significantly from each other is reflected in multilevel reliability estimates. The logic of multilevel reliability is the somewhat analogous to that used in determining the reliability of mean scale score based on multiple items (i.e., coefficient alpha) except that the focus is on agreement among persons within the same group rather than agreement among items designed to measure the same scale. Coefficient alpha is based on the average correlation among items (agreement among the items) and the number of items. The reliability of a university mean response for the NSS is based on the extent of agreement among students within each university and the number of students responding from each university. The reliability of the group mean depends on the proportion of variance that is located between groups – measured by the intraclass correlation (ICC) – and the number of persons (n) in each group (Bliese, 2000; Goldstein, 2005; Ludtke, Marsh, Robitzsch, Trautwein, Asparouhov & Muthén, in press; Snijders & Bosker, 1999).

$$\text{Reliability} = \frac{n \cdot ICC}{1 + (n - 1) \cdot ICC} \quad (1)$$

Some authors (e.g., Bliese, 2000) distinguish between ICC(1) –equivalent to ICC in equation 1—and ICC(2) which we refer to as the reliability of the group mean. In this sense, ICC(1) represents the assessment of single assessment (e.g., a score for a single person randomly selected from the group, sometimes referred to as single-rater reliability). Applying the typical Spearman-Brown formula (see equation 1), the researcher can estimate the reliability of the group mean, ICC(2), based on ICC(1) and the number of individuals within a group (n in equation 1).

Although the logic of this analogy between coefficient alpha and multilevel reliability is based on a two-level model, the approach is easily extended the three or more levels (e.g., Goldstein, 2005; Snijders & Bosker, 1999). For the three-level model considered here (L1=students, L2=disciplines within university, L3 = university) with corresponding variance components VC1, VC2, and VC3 (see table 7), the ICC1 for university the single-rater reliability is:

$$ICC1 = (VC3) / (VC1 + VC2 + VC3) \quad (2)$$

and the reliability for the university mean rating, ICC2, can be computed by applying equation 1. For discipline within university the single-rater reliability is

$$ICC1 = (VC2+VC3) / (VC1 + VC2 + VC3) \quad (3)$$

and the reliability for the discipline within group mean rating, ICC2, can be computed by applying equation 1.

In general: ICC1 for L3 is less than ICC1 for L2, but this need not be the case for ICC2 as the numbers of students at the university level will typically be much larger than numbers of students at discipline-within-university level. Using this approach it is also easy to estimate the minimum N required to achieve an "acceptable" reliability (e.g., .80) and a "good" reliability (e.g., .90) for mean ratings based on a particular university and a particular discipline-within-university group.

For 2005, the number of students in any one institution varied between 16 students to 13,706 ($M = 1223.71$), whilst for 2006, the number varied between 11 and 4871 ($M = 1092.85$) (see figure 4). Hence, it is not surprising that reliability estimates for different universities also vary substantially (see figure 5). In particular, there are some universities where the sample sizes are too small to provide a reliable estimate of the mean satisfaction. What is the minimum number of students at the university level that would provide acceptable levels of reliability for ratings of satisfaction at 0.90 and above? Rearranging the Spearman-Brown equation to find n , the minimum number of students in any given university would be 286 (for 2005) and 261 (for 2006). Inspection of Figures 4 and 5 demonstrates that most universities have N s substantially larger than this for both 2005 and 2006. Consequentially, for most universities, the reliability estimate is very high. For 2005, the 25th, 50th and 75th percentiles for estimates of reliability are .91, .97 and .98, whereas the corresponding values for 2006 are .93, .97, and .98. Nevertheless, there are some universities for which the sample sizes are sufficiently small to warrant caution in the interpretation of the results.

It is also important to evaluate the reliability of mean responses for discipline-within-university groups using the same procedure (described above). Since students within the same discipline-within-university group represent a more homogeneous group than students within from different disciplines, it follows that ICCs are higher. However, the number of students within each discipline-within-university group is highly variable and can be quite small. For example, in university 1 there are 1845 responses for Biological Sciences but only 46 for Creative Arts. Therefore, whilst the NSS responses could reliably differentiate between universities due to the large sample sizes, they may not be able to reliably differentiate between discipline-within-university groups due to the low numbers of students responding within each of these discipline groups. Even for the broadest categorisation of disciplines (with the smallest number of categories; 19), there are some disciplines that are not even offered at any particular university and many for which the number of students responding is small.

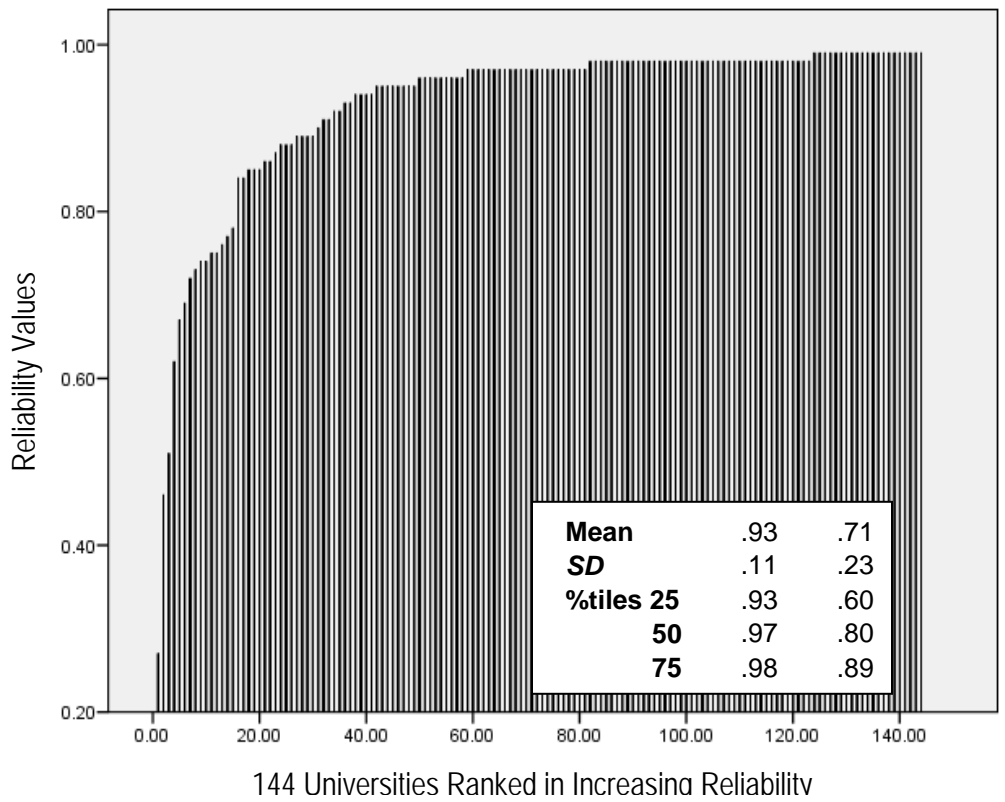
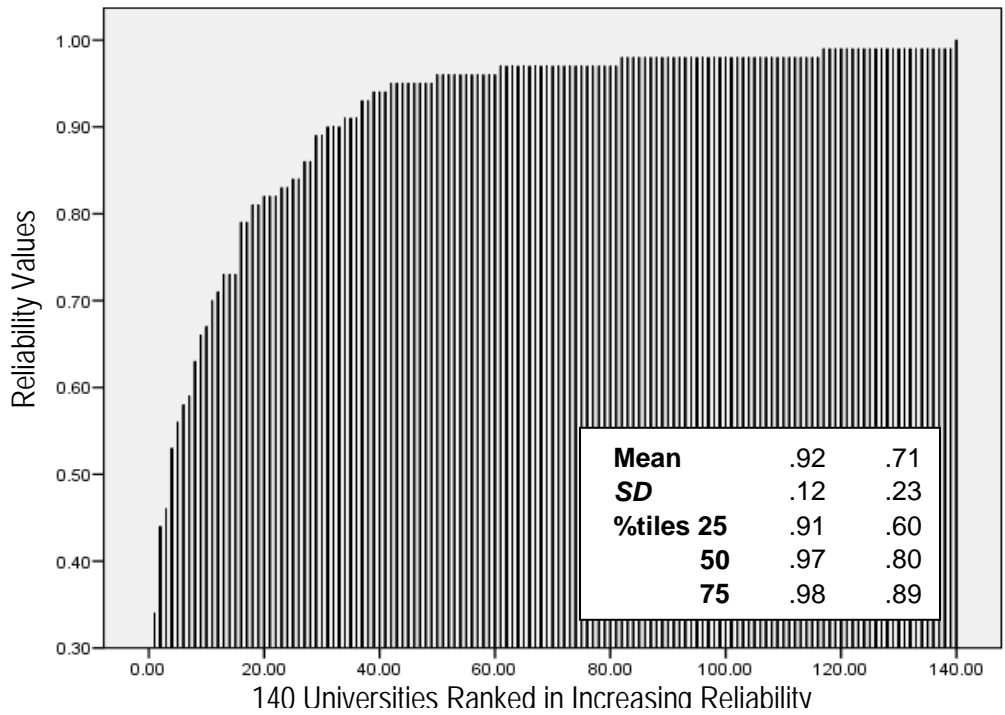


Figure 5. Distribution of reliability estimates across universities for 2005 (above) and 2006 (below) data

Frequency distributions of these Ns across discipline-within-university groups (Appendix 6A) demonstrate that the majority of these discipline-within-university groups are very small—at least in relation to providing reliable differentiation between groups. Also, the Ns vary substantially with the discipline classification scheme. Thus, for 2005 the median Ns are 68, 39, and 20 (for 19, 41, and 150 discipline categories respectively), whereas the median N for the 565-discipline classification in 2006 is only 5 students. In each case the number of students in each discipline group is very variable. Whilst some discipline-within-university groups achieve adequate levels of reliability (due in large part to larger sample sizes), the majority do not. What is the minimum number of students at the discipline-within-university group that would provide acceptable levels of reliability for ratings of satisfaction at 0.90 and above? Using the procedures described above, the minimum number of students would be 86 for 2005 and 98 for 2006 based on the broadest classification (of 19 disciplines) for both years.

When considering the reliability of responses at the discipline level, there is a “trade-off” between using more and less detailed discipline classifications. For 2005, discipline categories are broken down into 19, 41 and 150 disciplines and for 2006, they are categorised into 19, 41 and 565 disciplines. More detailed discipline breakdowns give better differentiation, but due to the smaller sample sizes of each group, the responses become less reliable. For the NSS results summarised here, the decreased reliability due to smaller Ns is more important than the increased differentiation due to more discipline categories. For both 2005 and 2006, the mean reliability across all discipline-within-university groups goes down as the number of discipline categories increase (see Appendix 6B). This poses an additional question of whether differences between ratings of the NSS occur as a result of the large sample size or whether the reliable differences between universities are sufficiently large to be substantively meaningful regardless of sample size. This has important practical implications, for example, in the construction of error bars around the mean of each university and for each discipline-within-university group within a university.

How stable are the rank ordering of universities from 2005 and 2006?

Stability over time is an alternative perspective from which to evaluate the NSS responses. Nearly all of the 145 universities considered here participated in the NSS for both 2005 and 2006. At the level of the university, the rankings are paired even though the students participating in the two data collection exercises were completely different. Nevertheless, given the substantial reliability of the ratings in each year considered separately and the fact that many of the 2006 participants were younger classmates of the 2005 participants, it would be expected that rankings of universities for the two years would be substantial. Consistent with expectations, based on those universities participating in both years, there is a very strong association between the ranking of a university in 2005 and 2006 ($r = 0.86$, $p < 0.01$) which is illustrated clearly by the scatterplot in figure 6 (also see Appendix 6 for a listing of the actual values for each university in 2005 and 2006). The overall satisfaction levels were also very stable across the two years ($Mn = 3.98$ in 2005, 4.00 in 2006). In summary, overall satisfaction ratings were highly stable over the two years considered.

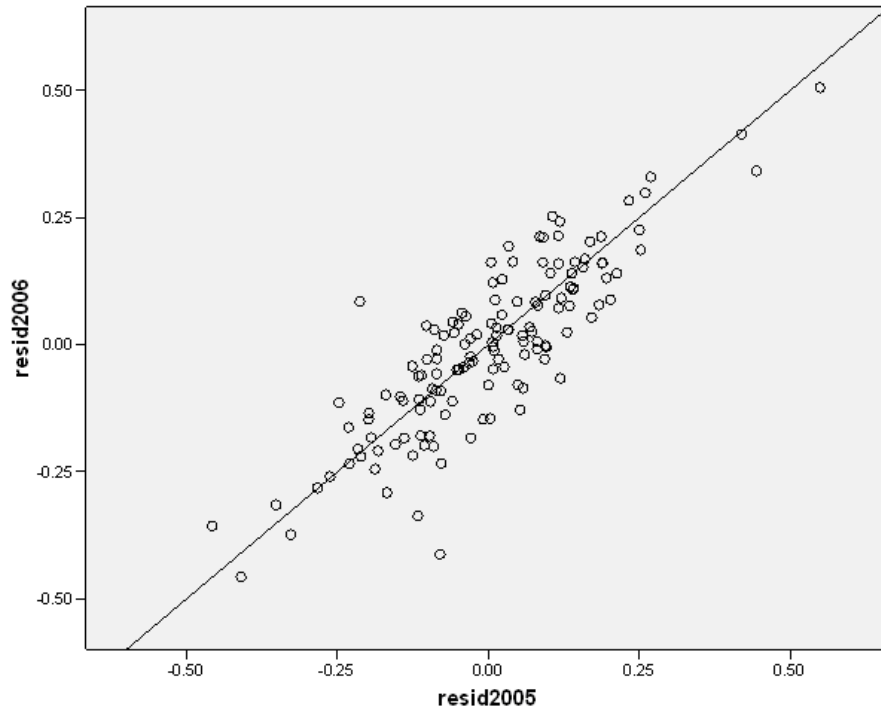


Figure 6. Scatter diagram of relations between ranking of universities based on the Overall Satisfaction rating for 2005 and 2006 (see Appendix 5 for the actual values for each university). The correlation between 2005 and 2006 ratings is .86

6 SUMMARY & DISCUSSION

Background

The purpose of this interim report is to provide an overview of preliminary results based on the 2005 and 2006 responses to the National Student Survey (NSS). All final year undergraduates in all UK universities are asked to complete the NSS about their educational experience. Consistent with the purpose of benchmarking UK universities, the main aims of the NSS are to: (1) help inform the choices of prospective students, 2) contribute to public accountability, and 3) provide useful data to institutions to use in their enhancement activities. Particularly in relation to the first two aims, the minimum condition for evaluating the appropriateness of the NSS responses is that they are able to clearly discriminate between UK universities and discipline-within-university groups in terms of student satisfaction with their undergraduate experience.

There is widespread use of students' evaluations of teaching effectiveness and reviews of this research based on thousands of publications (e.g., Marsh & Dunkin, 1997; Marsh & Roche, 1993; 1997; Marsh, 1987, 2007) have consistently shown that, with careful attention to measurement and theoretical issues, students' evaluations of teaching are: 1) multidimensional; 2) reliable and stable; 3) primarily a function of the instructor who teaches a course rather than the course that is taught; 4) relatively valid against a variety of indicators of effective teaching; 5) relatively unaffected by a variety of variables hypothesised as potential biases, such as expected course grades, class size, workload and prior subject interest; and 6) demonstrably useful in improving teaching effectiveness when coupled with concrete enhancement strategies in specific areas that teachers target for improvement.

Despite the very large literature on the use of students' evaluations to evaluate the teaching effectiveness of individual university teachers in particular classes, there is surprising little research on the use of responses by students to evaluate entire universities or discipline-within-university groups. Based on Australian research of PhD student's evaluations of their research experience, Marsh, Rowe and Martin (2002) found that the responses had good psychometric properties (reliability and factor structure) when evaluated at the level of the individual student. However, they argued that the appropriate unit of analysis for purposes of benchmarking should be either the university or the discipline-within-university group. When they evaluated PhD student responses with more appropriate multilevel analyses, they found that there were no statistically significant differences between the Australian universities or between discipline-within-university groups. On this basis they argued that the ratings were completely unreliable for the benchmarking purposes for which they were designed, and should not be used for this purpose. This previous research and methodological approach formed the basis for analyses conducted here.

NSS Factor Structure

We began our study with a careful evaluation of the factor structure underlying the responses to the 22 NSS items. Although the NSS is designed to measure six factors (Teaching; Assessment & Feedback; Support; Organisation; Resources; Personal Development; see Appendix 1) – plus an overall satisfaction rating – there has been surprisingly little research actually testing this factor structure. Based on a combination of exploratory and confirmatory factor analyses using both 2005 and 2006 data, we concluded that the most appropriate factor structure had seven specific factors (the Feedback/Assessment factor was divided into two factors) and an overall satisfaction factor (based on the response to the overall satisfaction item). This best fitting factor structure was highly consistent across 2005 and 2006 responses as demonstrated by confirmatory factor analysis tests of invariance across the two year groups.

We also evaluated higher-order factor analyses in which a single global satisfaction factor was posited to explain relations among the first-order factors. Strong support for this hierarchical model might support using a single summary score to summarise the NSS responses. Although the fit of this model was very good, it was not as good as the corresponding first-order models. Because the contributions of the different first-order factors to the higher-order factor differed substantially, the results argue against the use of an unweighted average of NSS responses (factors or items). Also, there were substantial amounts of variance in ratings in many of the first-order factors that were not explained by the single global satisfaction factor. However, further research is needed to explore further the practical implications for ignoring variance specific to particular factors – particularly in relation to providing information to universities and discipline-within-university groups that is useful for enhancing the quality of the undergraduate education experience.

The first-order and higher-order factors are both based on all 22 items. Either the higher-order factor score (a weighted average of the 7 specific factors and overall rating item) *or* the overall rating item by itself would be a reasonable single score summary, but information about the specific factors is lost by doing this. Since the specific factors are not equally important in terms of their contribution to the overall summary (either summary rating item or higher-order factor) it is probably *not* appropriate to take a simple unweighted average of the scale scores or the items to obtain a single summary score.

Differentiation Between Universities and Discipline-Within-University Groups

The main focus of the present investigation was to evaluate the ability of NSS responses to differentiate between universities and between discipline-within-university groups. For purposes of this interim report, we focus specifically on responses to the overall satisfaction item. Based on the nature of the NSS data and previous research (Marsh, Rowe & Martin, 2002), it is argued here that the most appropriate analysis should be based on a multilevel model with three levels (level 1 = students, level 2 = discipline-within-university groups, level 3 = university). In different models we explored

the implications of controlling for differences in individual student characteristics and fixed effects of discipline categories.

Variance components at the university level were highly significant from a statistical perspective and highly reliable – due primarily to the very large sample sizes at nearly all universities. However, differences between universities explained only about 3% of the variance in individual student responses and this estimate of variance explained was further reduced (to about 2.5%) after controlling for discipline differences. Hence, there is much more variation in responses by students within each university than there is in responses between the different universities (i.e., there is substantial lack of agreement among students within each university in terms of their satisfaction with their overall educational experience). Nevertheless, because of the high reliability (due to large N s), the differences between universities were highly stable over the two years that we considered ($r = .86$). Interpretation of these results provides a dilemma. Whereas differences between universities explain only a small amount of the variance in the NSS responses, these very small differences between universities are highly reliable and stable over time. The critical question is whether these small differences between universities are sufficiently large to help inform the choices of prospective students—a primary purpose of the NSS.

Variance components at the level of discipline-within-university groups are more complex to interpret. Not surprisingly, more variance is explained by differences in discipline-within-university groups than by differences between universities. Within universities, there are systematic differences in the levels of satisfaction experienced by students in different disciplines. Furthermore, these differences between discipline-within-university groups are larger when more detailed discipline classifications are considered (i.e., 41 or 150 discipline categories rather than 19 discipline categories). However, reliable differentiation between groups requires not only large differences between groups but also a sufficiently large number of students within each group so that these differences are reliable. Particularly for the more detailed discipline classifications, the number of students within each discipline-within-university group is too small to reliably differentiate between the groups. Further research is needed to achieve the optimal balance between the sample size (number of students within each discipline-within-university group) and the specificity of the disciplinary classification. On this basis we recommend that NSS ratings should only be used with appropriate caution to compare ratings by different discipline-within-university groups – either different disciplines within the same university or the same discipline across universities – and that any such results should be qualified in relation to interpretations of probable error based on appropriate multilevel models.

7 REFERENCES

- Abrami, P. C., d'Apollonia, S., & Cohen, P. A. (1990). Validity of student ratings of instruction: What we know and what we do not. *Journal of Educational Psychology, 82*, 219-231.
- Ainley, J., & Long, M. (1994). *The Course Experience Survey: The 1992 Graduates*. Graduate Careers Council of Australia, Department of Employment, Education and Training, Canberra, ACT.: Australian Government Printing Service.
- Bliese, P. D. (2000). Within-group agreement, non-independence, and reliability: Implications for data aggregation and analysis. In Klein, K. J., Kozlowski, S. W. J (Eds), *Multilevel theory, research, and methods in organizations: Foundations, extensions, and new directions*. US: Jossey-Bass.
- Bliese, P.D., & Hanges, P.J. (2004). Being both too liberal and too conservative: The perils of treating grouped data as though it is independent. *Organizational Research Methods, 7*, 400 – 417.
- Bligh, J., Lloyd-Jones, G., & Smith, G. (2000). Early effects of a new problem-based clinically oriented curriculum on students' perceptions of teaching. *Medical Education, 34*, 487 – 489.
- Bryk, A. S., Raudenbush, S. W. (1992). Hierarchical linear models: Applications and data analysis methods. *Hierarchical linear models: Applications and data analysis methods*. US: Sage Publications, Inc.
- Byrne, M., & Flood, B., (2003). Assessing the teaching quality of accounting programmes: An evaluation of the Course Experience Questionnaire. *Assessment & Evaluation in Higher Education, 28*(2), 135 – 145.
- Cashin, W. E. (1988). *Student Ratings of Teaching. A Summary of Research*. (IDEA paper No. 20). Kansas State University, Division of Continuing Education. (ERIC Document Reproduction Service No. ED 302 567).
- Cattell, R. B. (1966). The scree test for the number of factors. *Multivariate Behavioral Research, 1*, 245-276.
- Centra, J. A. (1993). *Reflective faculty evaluation*. San Francisco, CA: Jossey-Bass.
- Cohen, P. A. (1980). Effectiveness of student-rating feedback for improving college instruction: A meta-analysis. *Research in Higher Education, 13*, 321-341.
- d'Appollonia, S., & Abrami, P. C. (1997). Navigating student ratings of instruction. *American Psychologist, 52*, 1198–1208.

- Diseth, Å., Pallesen, S., Hovland, A., & Larsen, S. (2006). Course experience, approaches to learning and academic achievement. *Education & Training*, 48, 156 – 169.
- Ethington, C.A. (1996). A hierarchical linear modelling approach to studying college effects. In J.C. Smart (Ed.), *Higher education: Handbook of theory and research* (Vol. 12). New York: Agathon Press.
- Feldman, K. A. (1989a). Instructional effectiveness of college teachers as judged by teachers themselves, current and former students, colleagues, administrators, and external (neutral) observers. *Research in Higher Education*, 30, 137-194.
- Feldman, K. A. (1989b). Association between student ratings of specific instructional dimensions and student achievement: Refining and extending the synthesis of data from multisection validity studies. *Research in Higher Education*, 30, 583-645.
- Feldman, K. A. (1997). Identifying exemplary teachers and teaching: Evidence from student ratings. In R. P. Perry & J. C. Smart, (Eds.), *Effective Teaching in Higher Education: Research and Practice*. Agathon, New York, pp. 368–395.
- Feldman, K. A. (1998). Reflections on the effective study of college teaching and student ratings: one continuing quest and two unresolved issues. In J. C. Smart (Ed.) *Higher education: handbook of theory and research*. New York: Agathon Press, pp. 35-74.
- Gilmore, G. M., Kane, M. T., & Naccarato, R. W. (1978). The generalizability of student ratings of instruction: Estimates of teacher and course components. *Journal of Educational Measurement*, 15, 1-13.
- Goldstein, H. (1995). *Multilevel statistical models* (2nd ed.). London: Arnold.
- Greenwald, A. G., & Gillmore, G. M. (1997b). No Pain, No Gain? The importance of measuring course workload in student ratings of instruction. *Journal of Educational Psychology*, 89, 743-751.
- Higher Education Funding Council for England (HEFCE, October, 2005). Briefing notes for Students' Unions. Bristol, UK: HEFCE.
- Higher Education Statistics Agency (2007). http://www.hesa.ac.uk/index.php/component/option.com_studrec/task/show_file/Itemid,233/mnl,07051/href,a%5E%5EULN.html/ Date accessed: 1st December, 2007.
- Johnson, T. (1998). *The 1997 Course Experience Survey*. Graduate Careers Council of Australia, Department of Employment, Education and Training.

- Kaiser, H. F. (1960). The application of electronic computers to factor analysis. *Educational and Psychological Measurement*, 20, 141-151.
- L'Hommedieu, R., Menges, R. J., & Brinko, K. T. (1990). Methodological explanations for the modest effects of feedback. *Journal of Educational Psychology*, 82, 232-241.
- Lizzio, A., Wilson, K., & Hadaway, V. (2007). University students' perceptions of a fair learning environment: A social justice perspective. *Assessment & Evaluation in Higher Education*, 32(2), 195 – 213.
- Ludtke, O., Marsh, H. W., Robitzsch, A., Trautwein, U., Asparouhov, T., & Muthén, B. (in press). The Multilevel Latent Covariate Model: A New, More Reliable Approach to Group-Level Effects in Contextual Studies. *Psychological Methods*.
<<http://www.cmm.bristol.ac.uk/research/Lemma/index.shtml>>
<<http://www.statmodel.com/download/Ludtkeposted.pdf>>
- Luke, D.A. (2004). *Multilevel Modeling*. London: Sage Publications, Inc.
- Marsh, H. W. (1984). Students' evaluations of university teaching: Dimensionality, reliability, validity, potential biases, and utility. *Journal of Educational Psychology*, 76, 707-754.
- Marsh, H. W. (1987). Students' evaluations of university teaching: Research findings, methodological issues, and directions for future research. *International Journal of Educational Research*, 11, 253-388. (Whole Issue).
- Marsh, H. W. (2007). Students' evaluations of university teaching: A multidimensional perspective. In R. P. Perry & J C. Smart (Ed.), *The Scholarship of Teaching and Learning in Higher Education: An Evidence-Based Perspective*. New York: Springer.
- Marsh, H. W., & Bailey, M. (1993). Multidimensionality of students' evaluations of teaching effectiveness: A profile analysis. *Journal of Higher Education*, 64, 1-18.
- Marsh, H. W. & Dunkin, M. (1997). Students' evaluations of university teaching: A multidimensional perspective. In R.P. Perry & J.C. Smart (eds.), *Effective Teaching in Higher education: Research and Practice*. (pp. 241-320). New York: Agathon.
- Marsh, H. W. & Hattie, J. (2002). The relationship between research productivity and teaching effectiveness: Complimentary, antagonistic or independent constructs. *Journal of Higher Education*, 73, 603-642.

- Marsh, H. W., & Roche, L. A. (2000). Effects of grading leniency and low workloads on students' evaluations of teaching: Popular myth, bias, validity or innocent bystanders? *Journal of Educational Psychology*, 92, 202-228.
- Marsh, H. W., & Roche, L.A. (1997). Making students' evaluations of teaching effectiveness effective. *American Psychologist*, 52, 1187-1197.
- Marsh, H. W., & Roche, L.A. (1993). The use of students' evaluations and an individually structured intervention to enhance university teaching effectiveness. *American Educational Research Journal*, 30, 217-251.
- Marsh, H. W., Rowe, K., Martin, A. (2002). PhD students' evaluations of research supervision: Issues, complexities and challenges in a nationwide Australian experiment in benchmarking universities. *Journal of Higher Education*, 73 (3), 313-348.
- McKeachie, W. J. (1997). Student Ratings: The Validity of Use. *American Psychologist*, 52, 1218-25.
- McKeachie, W.J. (1979). Student ratings of faculty: A reprise. *Academe*, 65, 384- 397.
- McKinnon, K. R., Walker, S. H., Davis, D. (2000). *Benchmarking: A manual for Australian universities*. Canberra: Australian Department of Education, Training and Youth Affairs.
- Monk, D.H. (1992). Education productivity research: An update and assessment of its role in education finance reform. *Education Evaluation and Policy Analysis*, 14, 307 – 332.
- Ramsden, P. (1991). A performance indicator of teaching quality in higher education: The Course Experience Questionnaire. *Studies in Higher Education*, 16, 129-150.
- Richardson, J. (2005). National student survey: Interim assessment of the 2005 questionnaire. A report to the Higher Education Funding Council, 1 – 31.
- Richardson, J. (2006). Approaches to studying and perceptions of academic quality in a short web-based course. *British Journal of Educational Technology*, 34, 433 – 442.
- Sinharay, S., Stern, H.S., & Russell, D. (2001). The use of multiple imputation for the analysis of missing data. *Psychological Methods*, 6(4), 317 – 329.
- Snijders, T., Bosker, R. (1999). *Multilevel analysis: An introduction to basic and advanced multilevel modeling*. US: Sage Publications, Inc.

Surridge, P. (2006). *The National Student Survey 2005: Findings*. A report to the Higher Education Funding Council, 1 – 127.

Surridge, P. (2007). *The National Student Survey 2006: Findings*. A report to the Higher Education Funding Council, 1 – 132.

APPENDIX 1 - THE 22 NSS ITEMS BROKEN DOWN INTO 8 FACTORS

Teaching

1. Staff are good in explaining things.
2. Staff have made the subject interesting.
3. Staff are enthusiastic about what they are teaching
4. The course is intellectually stimulating

Assessment Fairness²

5. The criteria used in marking have been made clear in advance
6. Assessment arrangements and marking have been fair

Assessment Feedback²

7. Feedback on my work has been prompt
8. I have received detailed comments on my work
9. Feedback on my work has helped me clarify things I did not understand

Support

10. I have received sufficient advice and support with my studies
11. I have been able to contact staff when I needed to
12. Good advice was available when I needed to make study choices

Organisation

13. The timetable works efficiently as far as my activities are concerned
14. Any changes in the course or teaching have been communicated effectively
15. The course is well organised and is running smoothly

Resources

16. The library resources and services are good enough for my needs
17. I have been able to access general IT resources when I needed to
18. I have been able to access specialised equipment, facilities or rooms when I needed to

Personal Development

19. The course has helped me to present myself with confidence
20. My communication skills have improved
21. As a result of the course, I feel confident in tackling unfamiliar problems

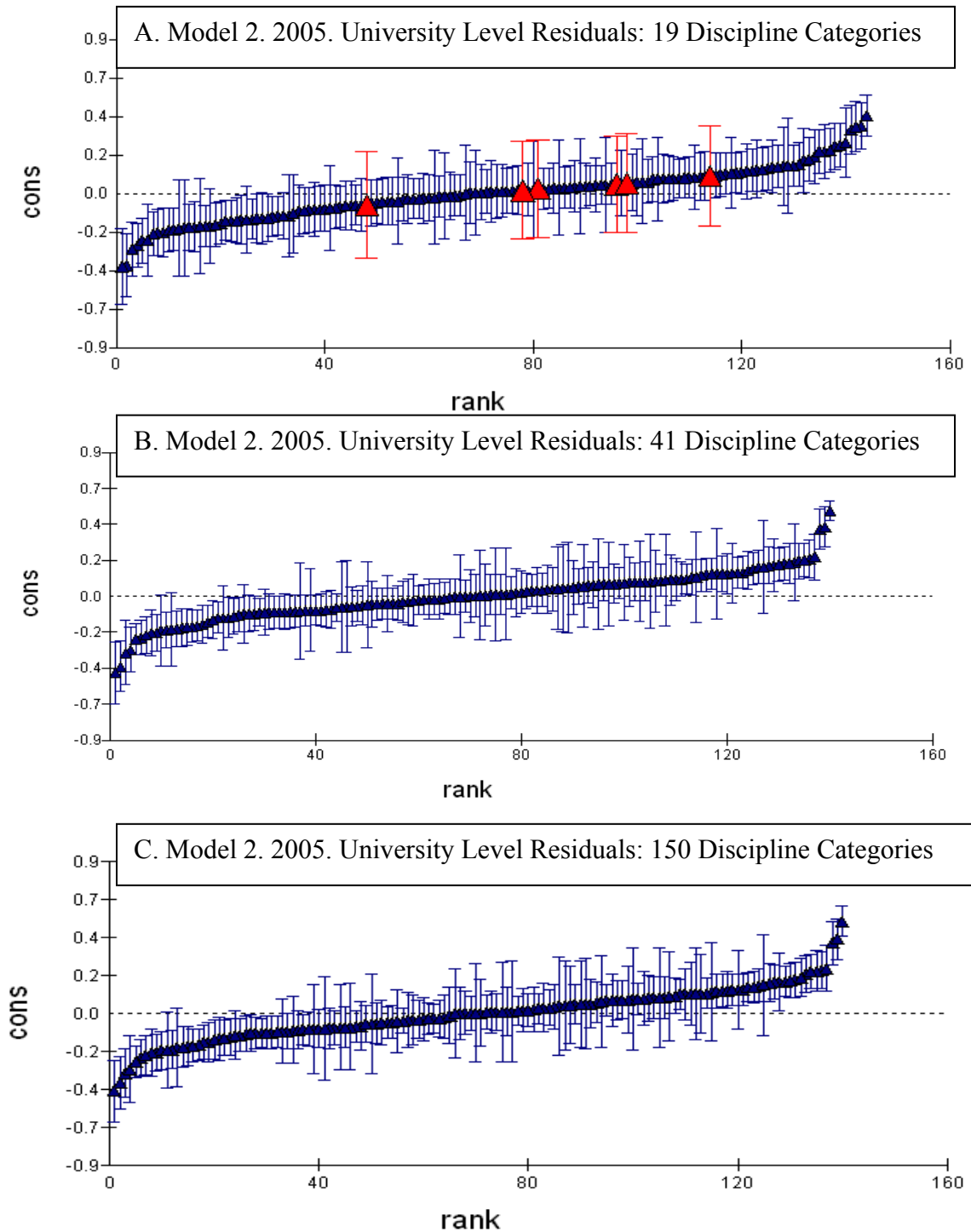
Overall Satisfaction

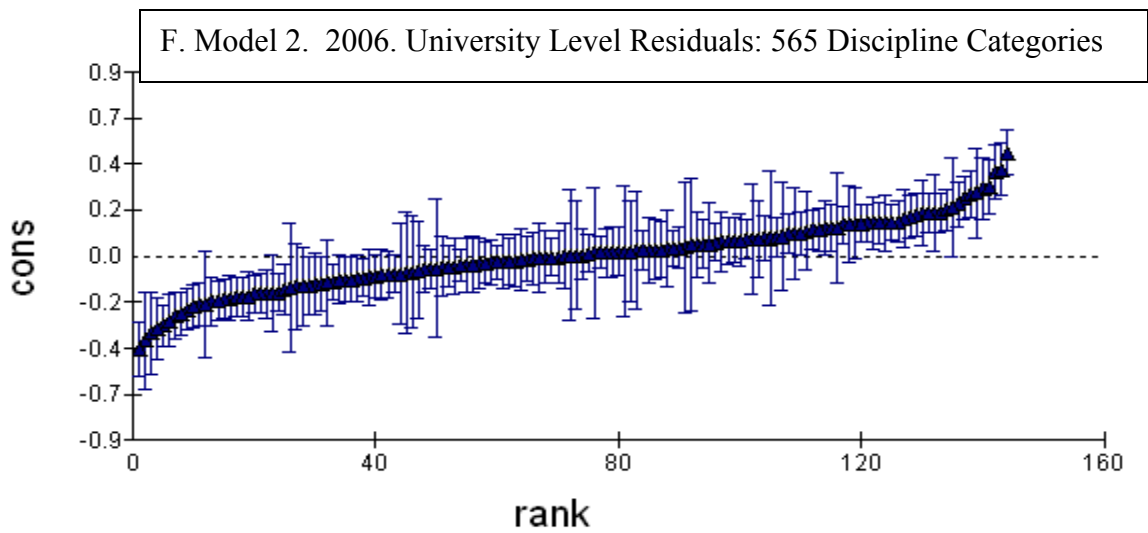
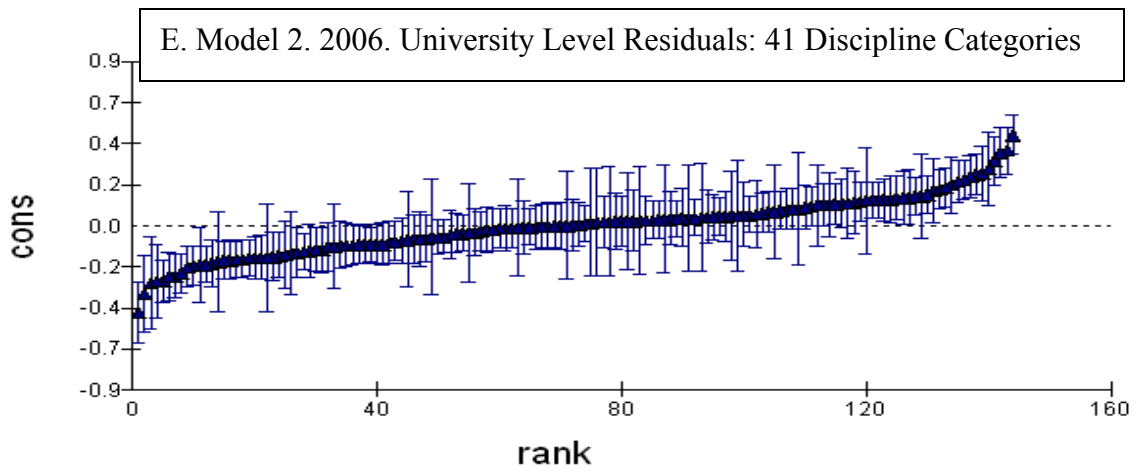
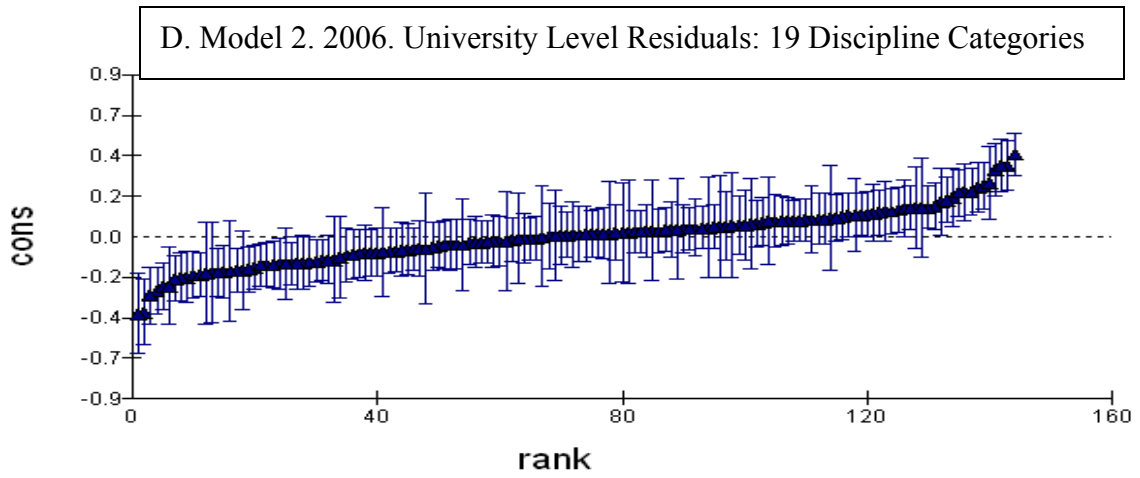
22. Overall, I am satisfied with the quality of the course.

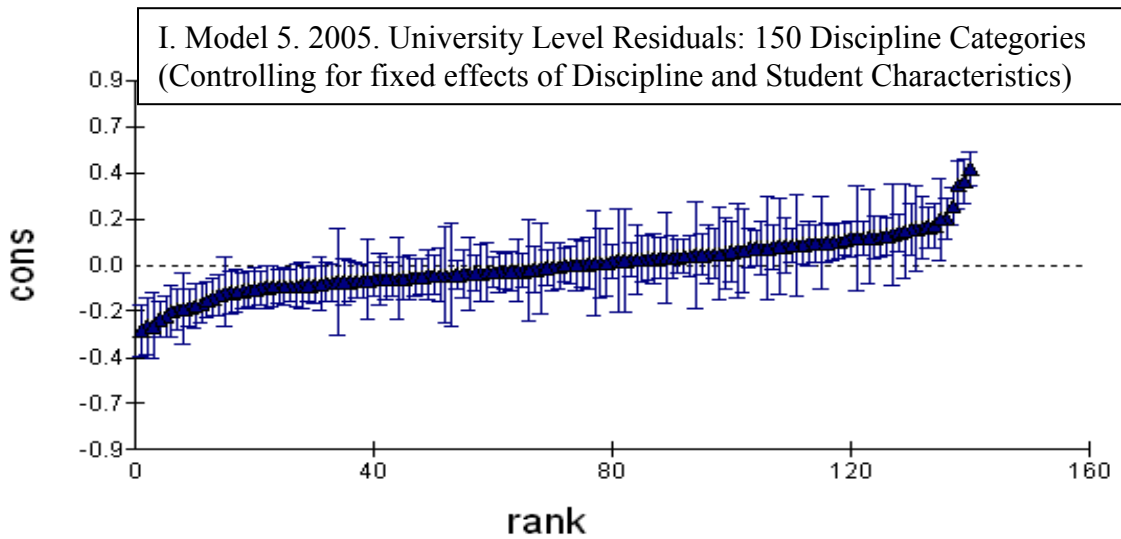
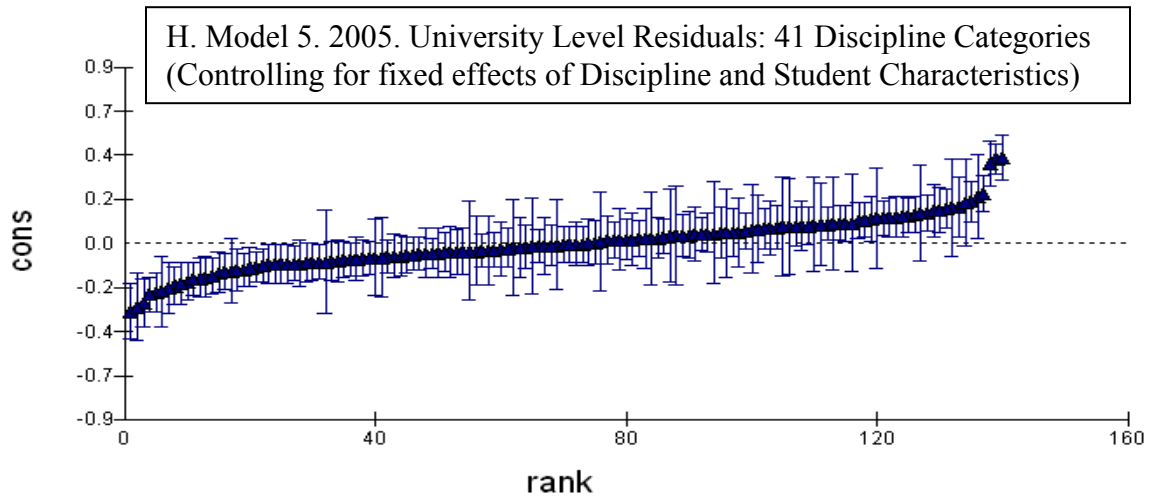
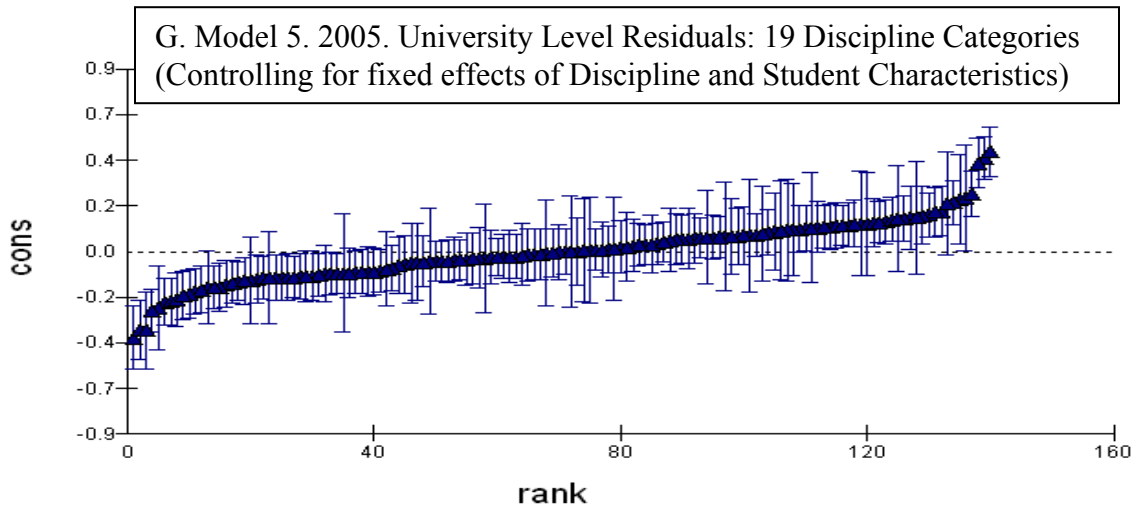
² According to the a priori design of the NSS survey, these two factors were combined to form an overall assessment factor, but results presented here suggest that they should be separated to identify distinct components of assessment.

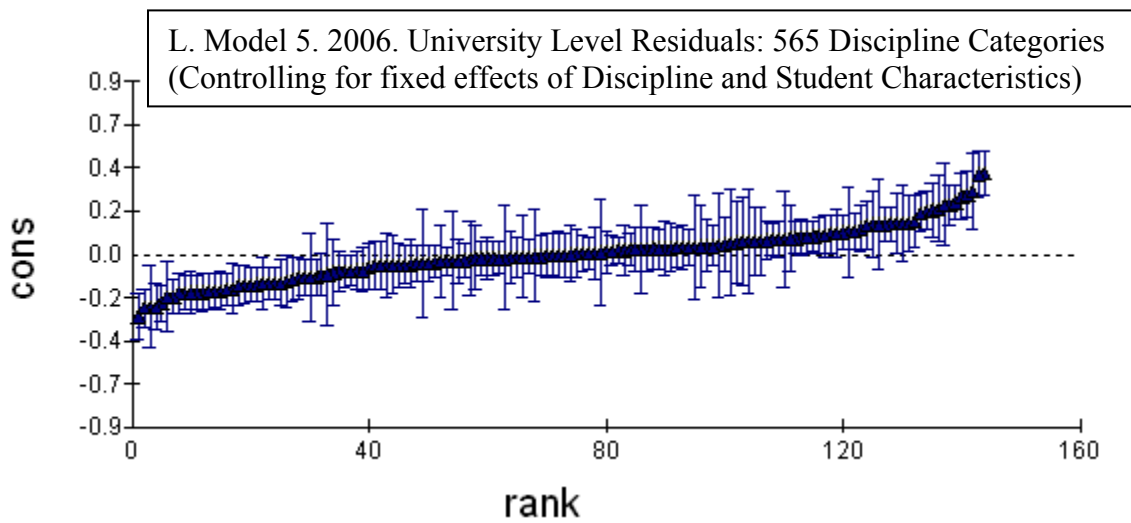
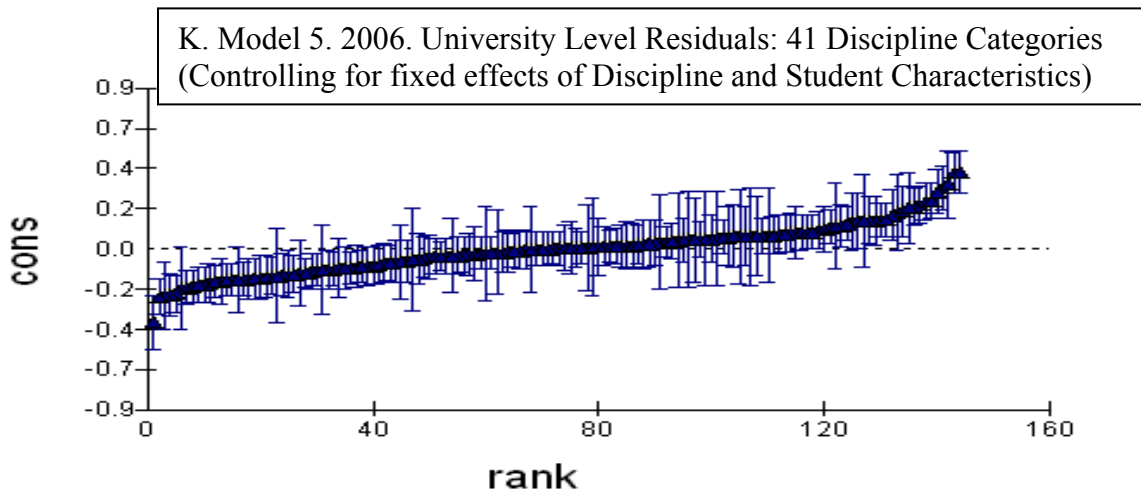
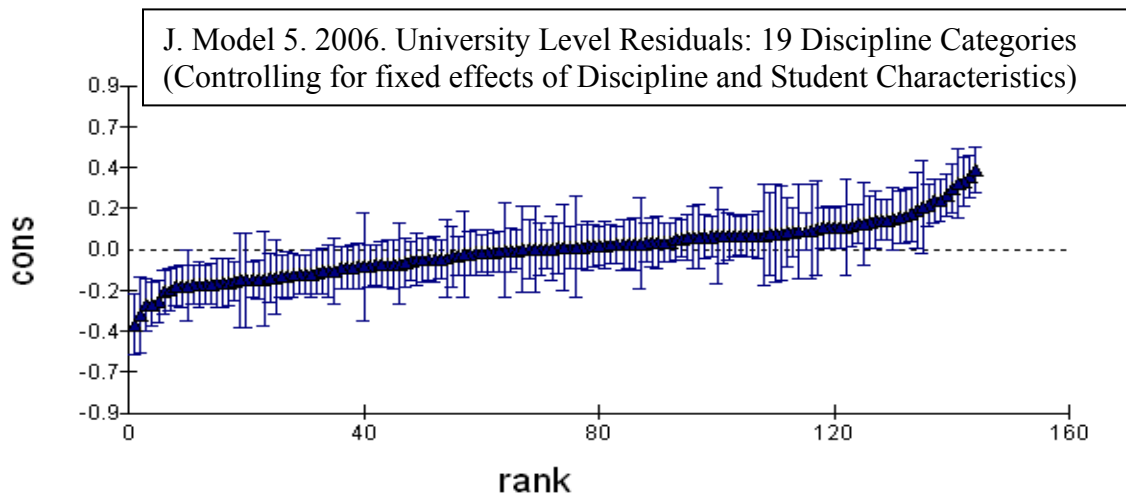
APPENDIX 2 – CATERPILLAR PLOTS OF UNIVERSITIES

Caterpillar Plots of Universities for Models 2 & 5 in Table 7: Mean and Probable Error for Each University Ranked From Lowest to Highest (dotted line at zero is the grand mean across all universities)





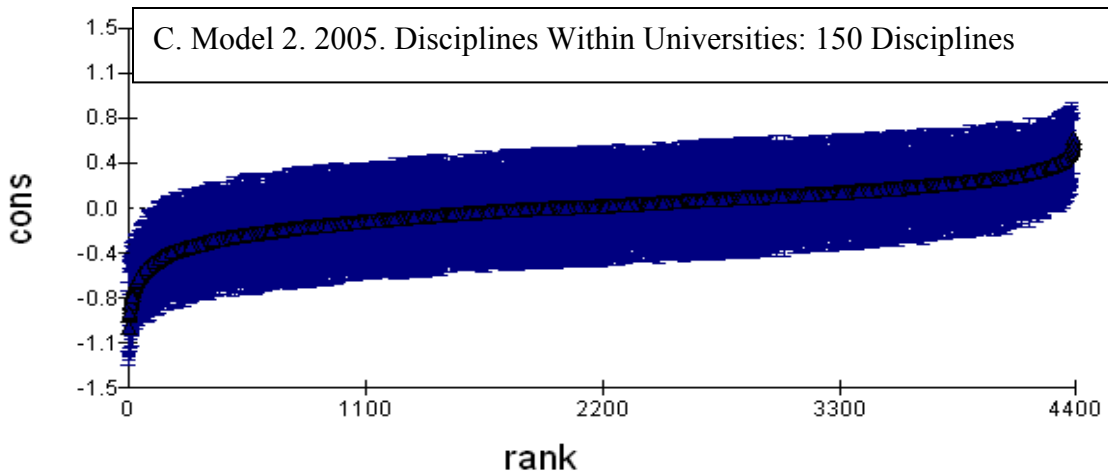
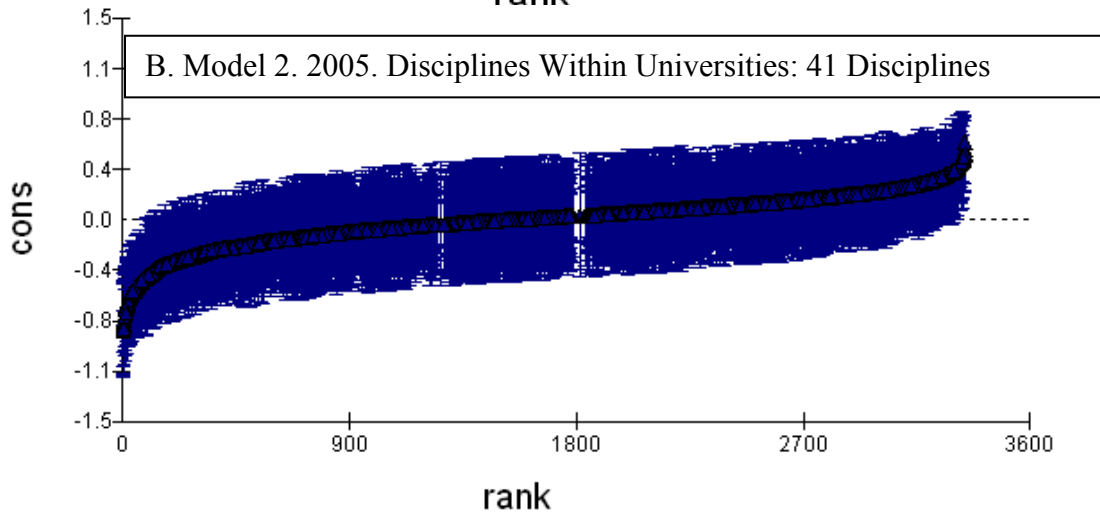
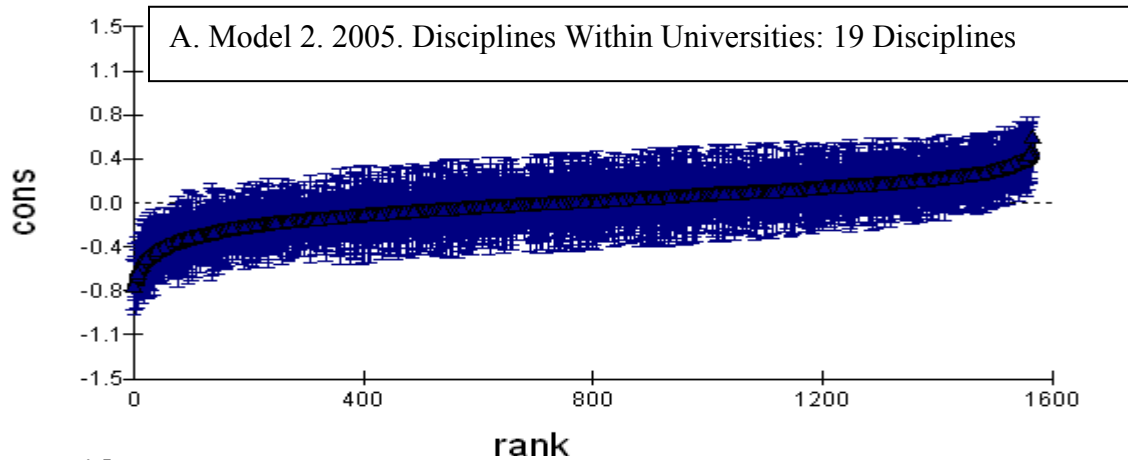


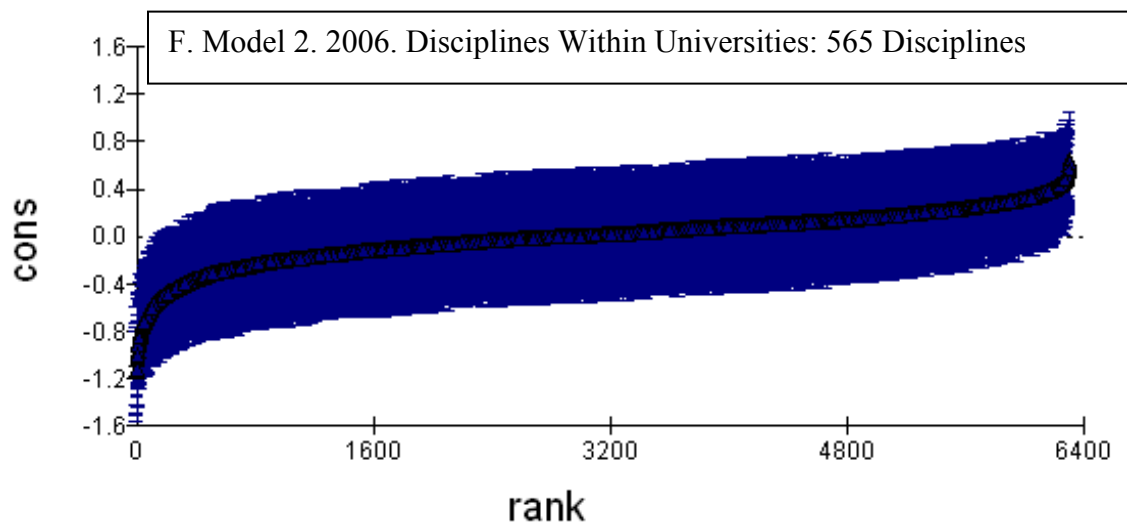
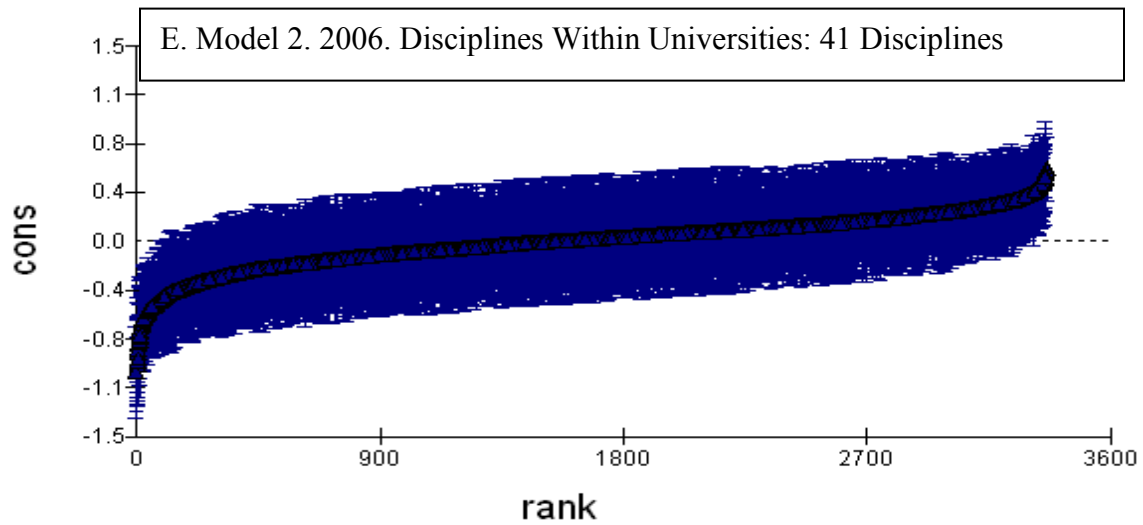
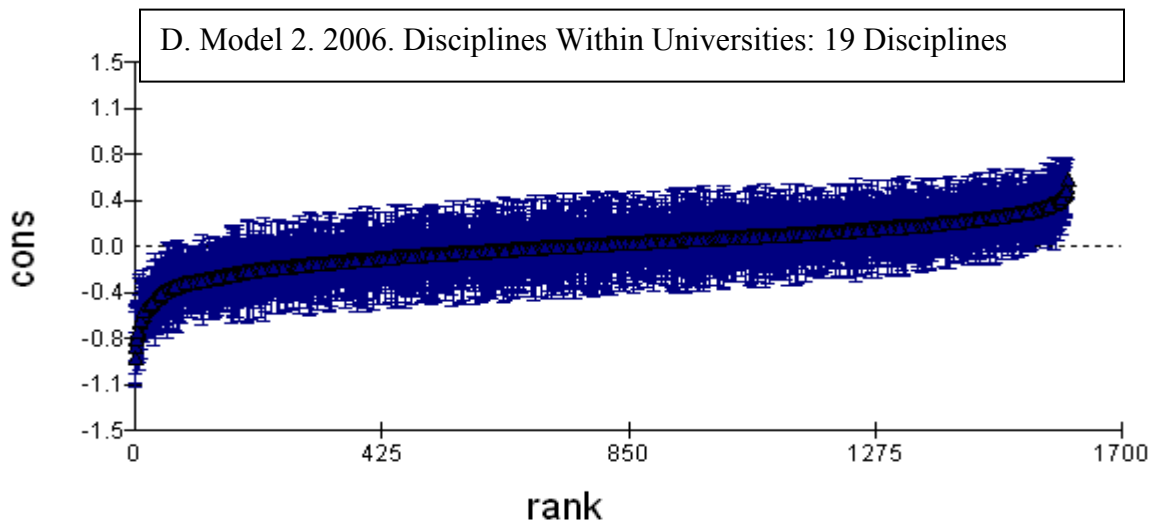


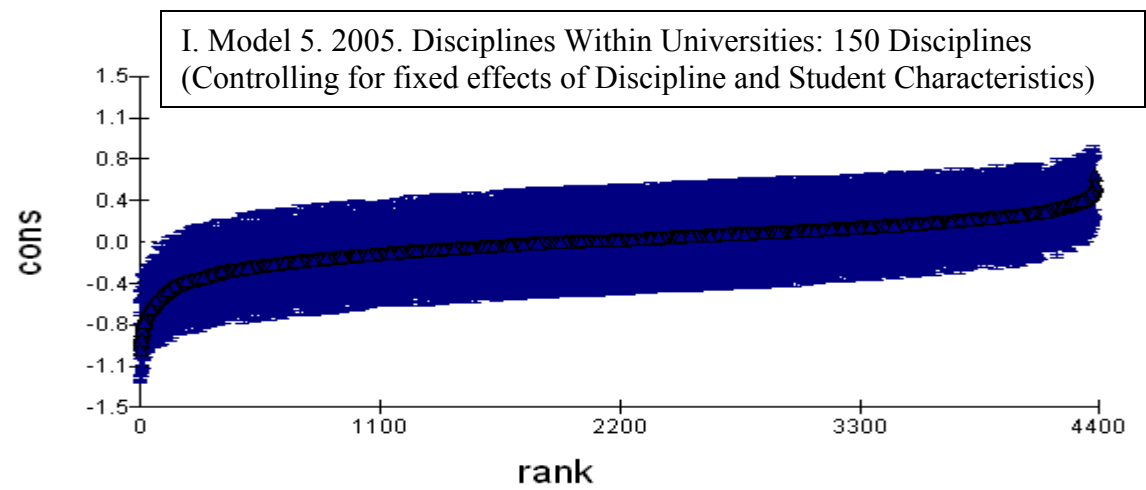
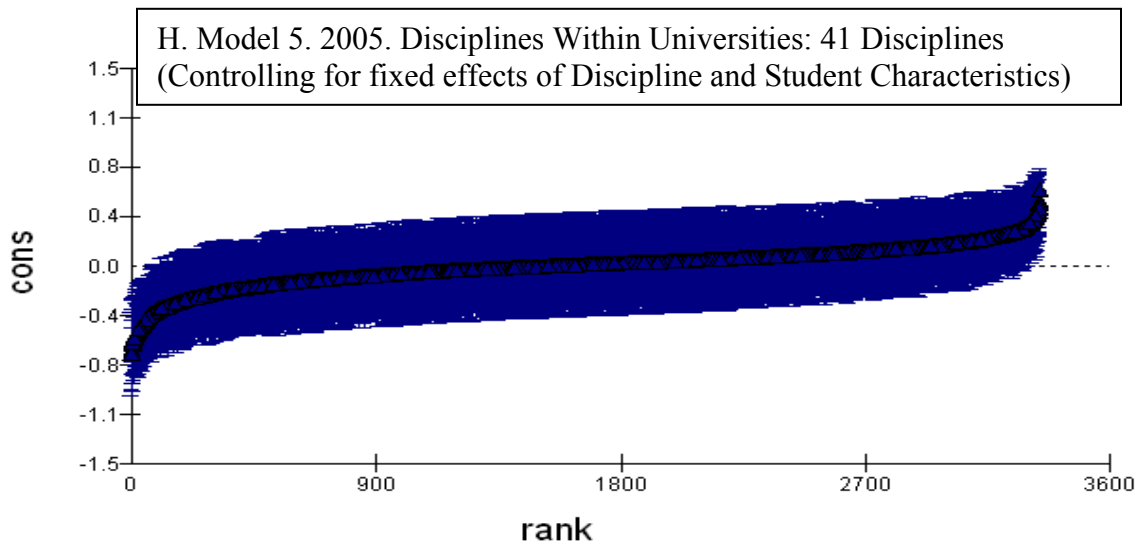
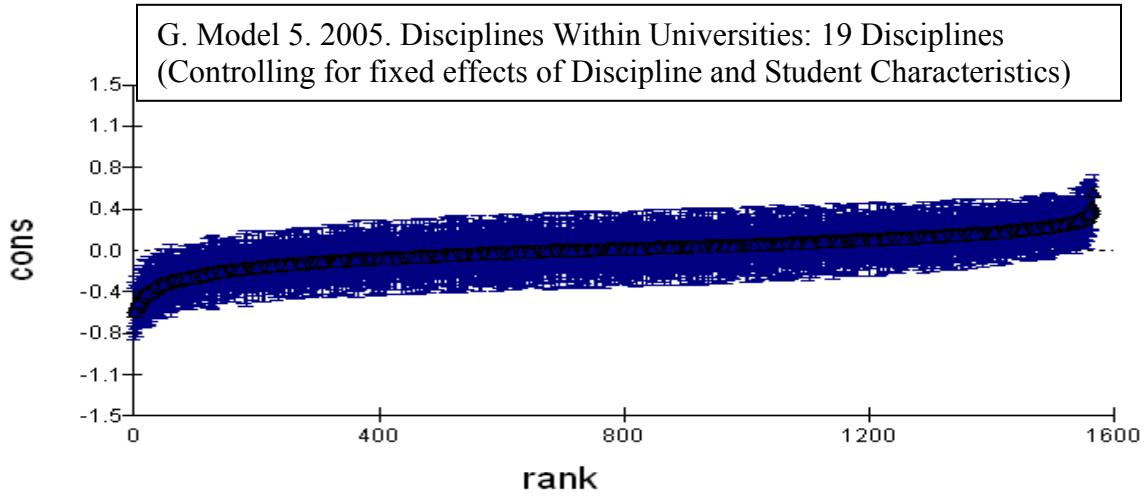
APPENDIX 3 – CATERPILLAR PLOTS OF DISCIPLINES WITHIN UNIVERSITIES

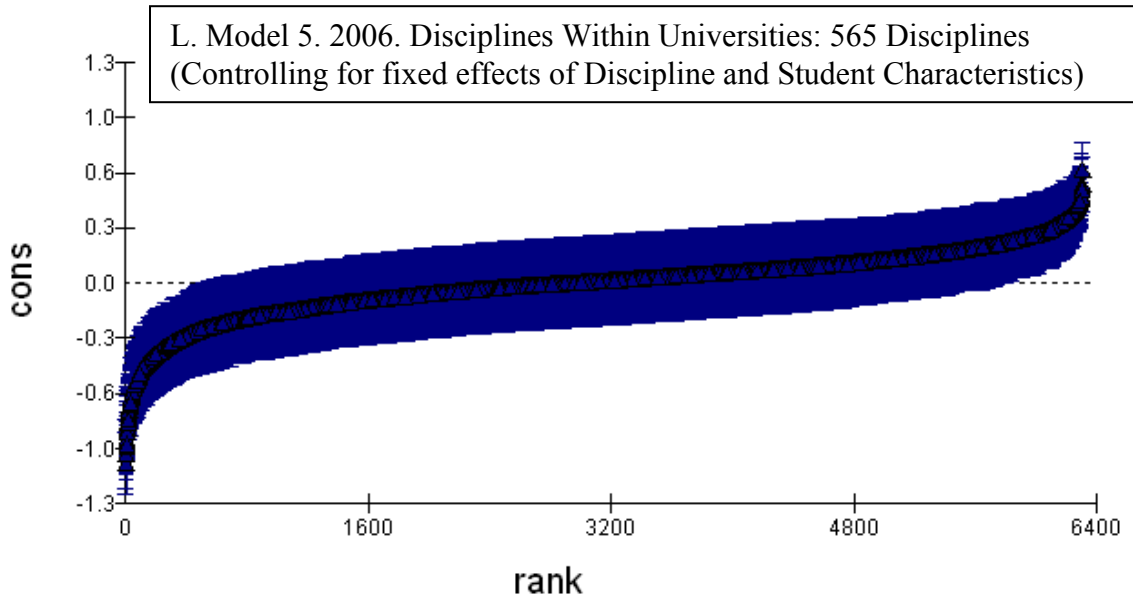
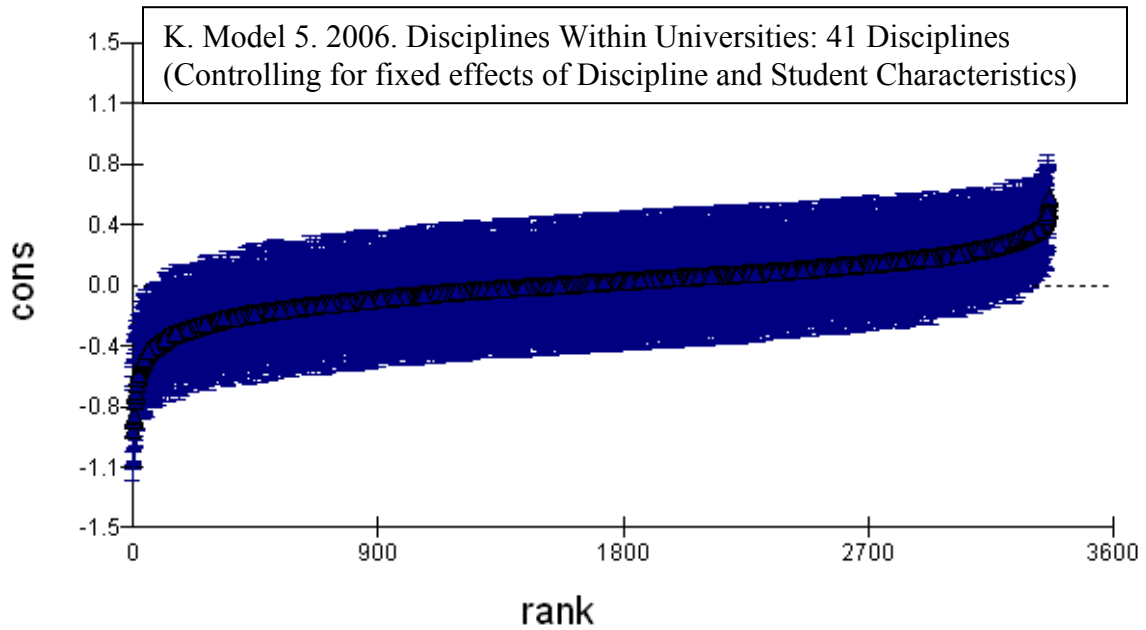
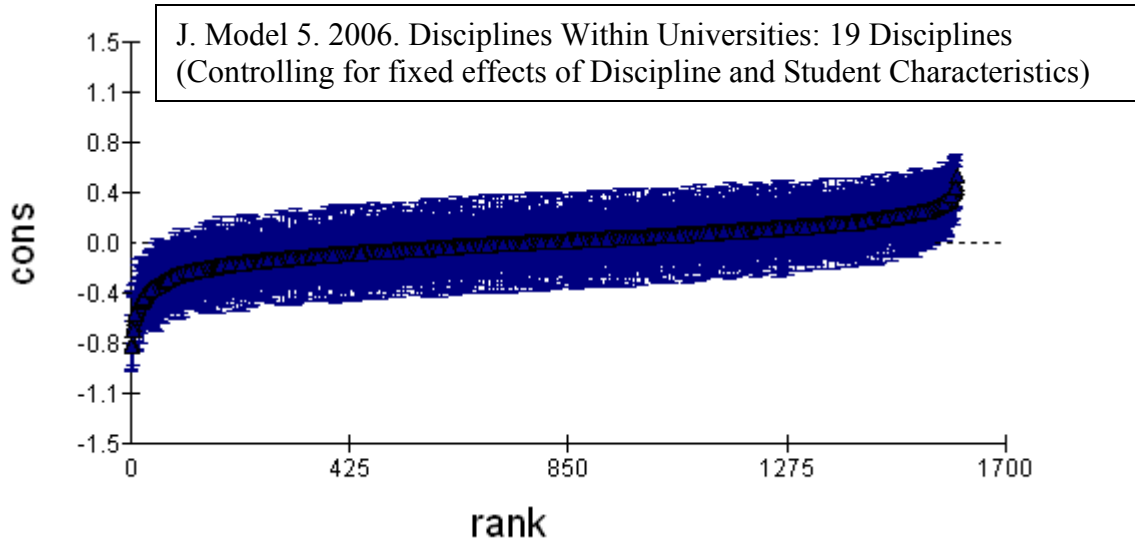
*For Models 2 & 5 in Table 7

* Mean and Probable Error for Each Discipline-within-University group Ranked From Lowest to Highest (dotted line at zero is the grand mean across all groups)









APPENDIX 4 – FIXED EFFECTS ASSOCIATION WITH DUMMY VARIABLES OF EACH OF THE DISCIPLINE CLASSIFICATIONS USED IN 2005 AND 2006

*based on discipline classification with 19 categories

2005 Data	Coefficient (std. error) with discipline as fixed effect	Discipline + individual student characteristics
(Medicine & Dentistry)		
Subjects Allied to Medicine	0.052(0.051)	0.030(0.051)
Biological Sciences	0.040(0.049)	0.024(0.049)
Veterinary Sciences, Agriculture, and related subjects	-0.040(0.059)	-0.060(0.059)
Physical Sciences	0.136(0.051)	0.122(0.051)
Mathematical and Computer Sciences	-0.113(0.049)	-0.123(0.049)
Engineering	-0.100(0.051)	-0.114(0.051)
Technologies	-0.166(0.062)	-0.185(0.062)
Architecture, Building and Planning	-0.138(0.055)	-0.167(0.055)
Social studies	-0.001(0.049)	-0.022(0.049)
Law	0.054(0.051)	0.039(0.051)
Business and Administrative Studies	-0.057(0.049)	-0.075(0.049)
Mass Communications and Documentation	-0.125(0.052)	-0.145(0.052)
Linguistics, Classics and related subjects	0.201(0.050)	0.171(0.050)
European Languages, Literature and related subjects	0.044(0.056)	0.022(0.056)
Eastern, Asiatic, African, American and Australasian Languages	0.051(0.064)	0.035(0.065)
Historical and Philosophical Studies	0.208(0.051)	0.174(0.051)
Creative Arts and Design	-0.189(0.049)	-0.212(0.050)
Education	-0.018(0.055)	-0.063(0.055)

Continued on the next page

2006 Data	Coefficient (std. error) with discipline as fixed effect	Discipline + individual student characteristics (std. error)
(Medicine & Dentistry)		
Subjects Allied to Medicine	0.118(0.051)	0.089(0.052)
Biological Sciences	0.083(0.049)	0.065(0.050)
Veterinary Sciences	-0.095(0.100)	-0.123(0.101)
Agriculture, and related subjects	-0.082(0.061)	-0.110(0.062)
Physical Sciences	0.154(0.051)	0.138(0.052)
Mathematical Sciences	0.098(0.054)	0.081(0.054)
Computer Science	-0.098(0.050)	-0.114(0.051)
Engineering and Technology	-0.050(0.051)	-0.068(0.051)
Architecture, Building and Planning	-0.083(0.055)	-0.120(0.056)
Social Studies	0.041(0.049)	0.019(0.050)
Law	0.134(0.051)	0.111(0.051)
Business and Administrative Studies	0.004(0.049)	0.013(0.050)
Mass Communications and Documentation	-0.080(0.052)	-0.103(0.053)
Languages	0.201(0.050)	0.177(0.050)
Historical and Philosophical Studies	0.256(0.051)	0.230(0.051)
Creative Arts and Design	-0.118(0.050)	-0.149(0.050)
Education	0.031(0.053)	0.012(0.053)
Combined	0.214(0.074)	0.161(0.075)

Note. For each year, two separate multilevel analyses were conducted: One that included only the 19 disciplined classifications as fixed effects (18 dummy variables with “Medicine & Dentistry” as the reference category) and one with both discipline classifications and student characteristics as fixed effects. Standard errors are presented in parentheses (effects greater than 2 standard errors are statistically significant and in red).

APPENDIX 5 - FIXED EFFECTS ASSOCIATED WITH DUMMY VARIABLES OF EACH OF THE STUDENT CHARACTERISTICS

*Each characteristic considered separately and in combination with all other student characteristics.

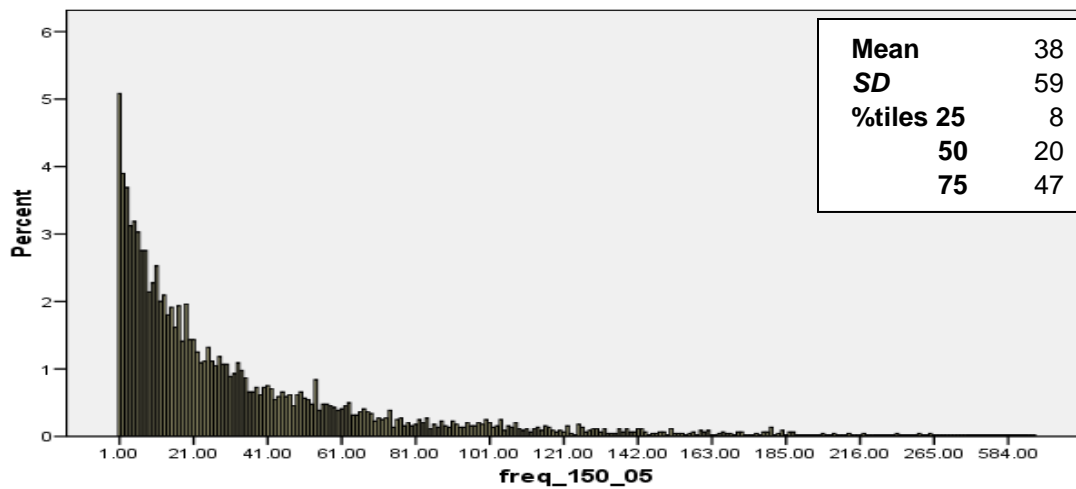
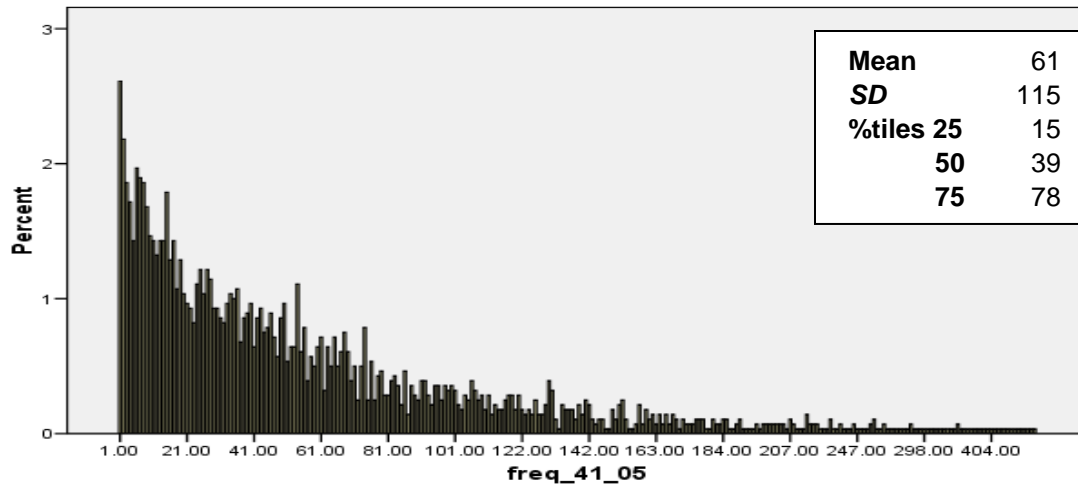
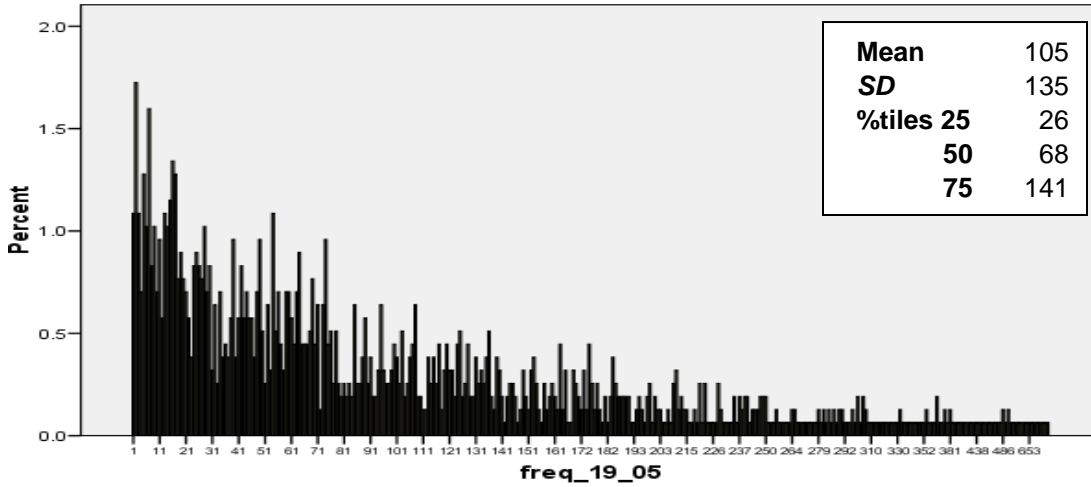
Student Characteristic		Coefficient (std. error) as a single fixed effect		Coefficient (std. error) when all effects are considered together	
		2005	2006	2005	2006
Gender (female); male	Male	-0.009(0.005)	-0.011(0.005)	-0.027(0.005)	-0.033(0.005)
Ethnicity (white)	Black	-0.014(0.013)	0.035(0.014)	-0.034(0.013)	0.016(0.014)
	Asian	-0.105(0.008)	-0.089(0.009)	-0.116(0.009)	-0.097(0.009)
	Other	-0.091(0.011)	-0.050(0.012)	-0.101(0.011)	-0.054(0.012)
	Unknown	-0.059(0.014)	-0.088(0.015)	-0.069(0.015)	-0.082(0.016)
Major source of tuition fees (no award)	Award	-0.003(0.005)	-0.002(0.005)	-0.003(0.005)	-0.003(0.005)
	Other	-0.079(0.039)	-0.012(0.032)	-0.095(0.040)	-0.033(0.032)
Age group (unknown for 2005); (18& under for 2006)	19	-0.041(0.006)	-0.026(0.006)	-0.035(0.006)	-0.022(0.006)
	20-21	-0.053(0.008)	-0.038(0.008)	-0.040(0.008)	-0.025(0.009)
	22-30	0.011(0.009)	-0.022(0.009)	-0.026(0.009)	-0.012(0.010)
	31-40	0.063(0.011)	0.032(0.012)	-0.082(0.012)	0.037(0.013)
	41+	0.074(0.012)	0.064(0.014)	0.105(0.014)	0.076(0.015)
Mode of study (full time)	Sandwich	0.040(0.011)	0.033(0.012)	0.052(0.012)	0.042(0.012)
	Part time	0.042(0.012)	0.036(0.018)	-0.001(0.014)	0.044(0.019)
	Other study	-0.047(0.178)	0.059(0.014)	-0.266(0.203)	0.003(0.015)
Domicile (UK)	Non-UK	-0.037(0.010)	-0.059(0.009)	0.028(0.011)	0.009(0.011)
Method of response (web)	Post	-0.016(0.007)	0.022(0.007)	-0.018(0.007)	0.020(0.007)
	Email/other	0.137(0.006)	0.187(0.006)	0.146(0.006)	0.195(0.006)
Accommodation	Parent	0.018(0.010)	0.032(0.010)	0.022(0.010)	0.020(0.010)

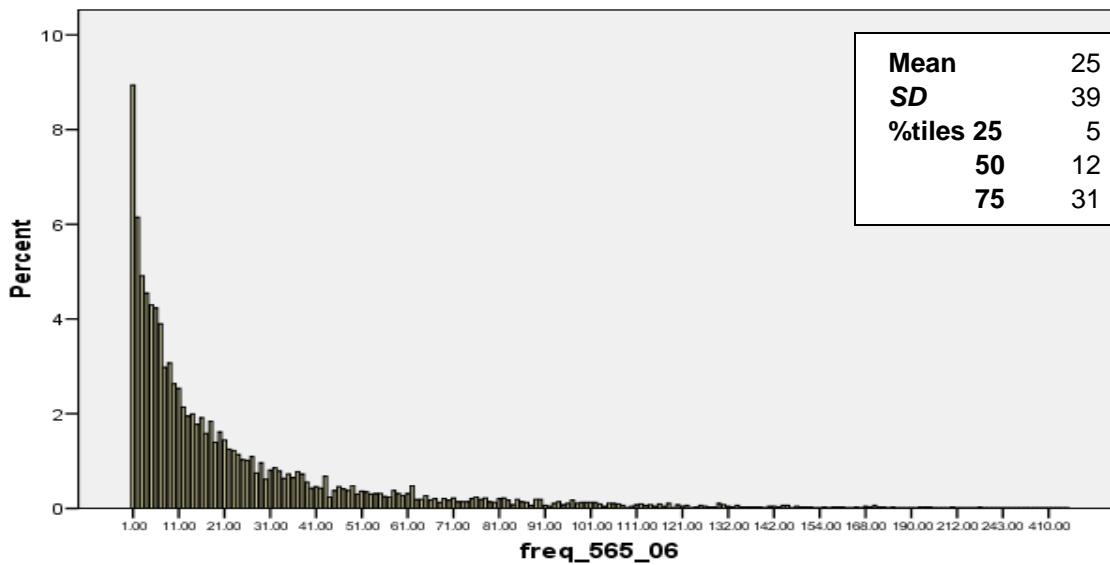
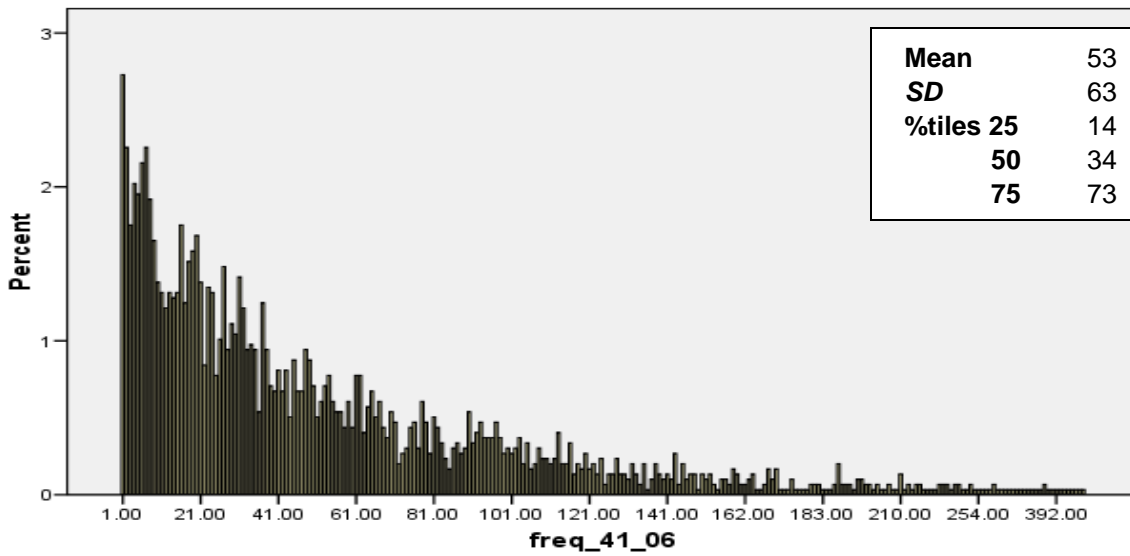
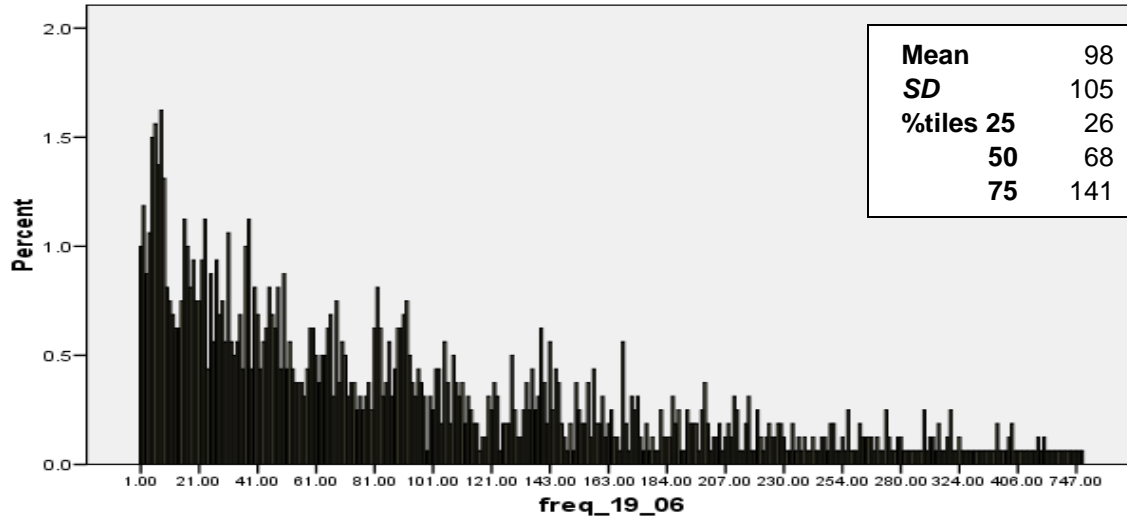
during term time (university)	Own home	0.023(0.009)	0.032(0.008)	0.002(0.009)	0.015(0.009)
	Unknown/other	0.009(0.010)	0.028(0.010)	-0.009(0.010)	0.005(0.010)
Disability (none)	Dyslexia	-0.052(0.013)	-0.075(0.013)	-0.060(0.013)	-0.083(0.013)
	Other	-0.008(0.012)	-0.056(0.013)	-0.012(0.012)	-0.061(0.013)
	Unknown	0.031(0.036)	0.008(0.029)	-0.024(0.038)	0.012(0.029)
Qualification Aim (first degree)	Other degree	0.001(0.009)	0.048(0.011)	-0.014(0.010)	-0.020(0.011)
UCAS score tariff (<200)	200-299	0.007(0.008)	0.002(0.008)	0.015(0.008)	0.006(0.008)
	300-399	0.011(0.008)	0.018(0.007)	0.014(0.008)	0.020(0.008)
	400-499	0.014(0.011)	0.039(0.009)	0.016(0.011)	0.038(0.010)
	500+	0.042(0.017)	0.042(0.014)	0.044(0.017)	0.042(0.014)
	Unknown	----	0.878(0.678)	----	0.923(0.674)

Note. For each year, two separate multilevel analyses were conducted: One that included each student characteristic considered separately and one that included all of the student characteristics. Results present here are based on a three level model (with level 2 = discipline with 41 discipline categories, but results were similar for other discipline classifications).

APPENDIX 6A - FREQUENCY DISTRIBUTION OF STUDENTS IN DIFFERENT “DISCIPLINE-WITHIN-UNIVERSITY” GROUPS FOR DIFFERENT DISCIPLINE CLASSIFICATIONS

* Classifications for 2005: 19, 41, & 150 disciplines; 2006: 19, 41, & 565 disciplines





**APPENDIX 6 B - FREQUENCY DISTRIBUTION OF RELIABILITY ESTIMATES
FOR THE AVERAGE OVERALL SATISFACTION RATING IN DIFFERENT
“DISCIPLINE-WITHIN-UNIVERSITY” GROUPS FOR DIFFERENT DISCIPLINE
CLASSIFICATIONS**

* Classifications for 2005: 19, 41, & 150 disciplines; 2006: 19, 41, & 565 disciplines

