Department of Education,
Training and Youth Affairs

Development of the Course Experience Questionnaire (CEQ)

Craig McInnis
Patrick Griffin
Richard James
Hamish Coates

Centre for the Study of Higher Education
and Assessment Research Centre

Faculty of Education
The University of Melbourne

01/1 April 2001

Evaluations and Investigations Programme
Higher Education Division

Instructions for pdf navigation

- Use the arrows on the Acrobat menu bar to navigate forwards or backwards page by page
- Alternatively, use the arrow icons on your keyboard to navigate through the document.
- To enlarge the viewing screen either:
  - use the magnifying glass by clicking on the area you wish to enlarge or by forming a marquee over the area you wish to view (ie. hold the mouse button down and drag the magnifying glass over the area); or
  - use the view options menu bar at the bottom of the Acrobat screen.
- To pan out from the page, hold down the option button on your keyboard to change the +ve symbol on the magnifying glass to a –ve symbol , then click the mouse.
- To search for a word or phrase use the binoculars icon on the menu bar.
- The Contents pages are live, ie. if you click on a topic you will go to that page.
- You can return to the Contents page by clicking your mouse on 'Contents' on the top of each page.

ISBN 0 642 45722 0

ISBN 0 642 45723 9 (Online version)

DETYA No. 6662 HERC01A

The report is funded under the Evaluations and Investigations Programme of the Department of Education, Training and Youth Affairs.

The views expressed in this report do not necessarily reflect the views of the Department of Education, Training and Youth Affairs.

Contents

Tables and figures

Tables

Figures

Acknowledgments

The project team is grateful for the support of the Advisory Committee: Mr Phil Aungles (DETYA); Mr Conor King (Australian Vice-Chancellors' Committee); Professor Michael Koder (Deputy Vice Chancellor, University of Sydney, and Chair of the GCCA Survey Management Group); Mr Gavin Moodie (Victoria University); and Professor Paul Ramsden (Pro-Vice-Chancellor, University of Sydney). The Committee provided detailed and invaluable advice on the methodology and scale development throughout the project.

Sue Griffin (ARC) and Rae Massey (CSHE) provided valuable assistance with the overall management of the project.

The project would not have proceeded without the practical support of staff from the participating universities. We are particularly grateful to the many academics and administrators who have gone out of their way to assist with the field testing programme.

Executive summary

This project was jointly conducted by the Centre for the Study of Higher Education (CSHE) and the Assessment Research Centre (ARC) of The University of Melbourne. The primary objective of the project was to prepare an extended form of the existing Course Experience Questionnaire (CEQ) to include measures of the broader aspects of the student experience while maintaining the integrity of the existing instrument. The impetus for enhancing the CEQ is recognition of the diversity of the Australian higher education system and the value attached to the full variety of educational and social experiences of students.

The major steps in the project were:

- Extensive consultation with the higher education community in preparing the conceptual framework guiding the development of new items and scales;
- Formation of an Advisory Committee and regular meetings to assist with the direction of the project;
- Drafting of potential new items and scales on the basis of feedback from the consultation process;
- Pilot testing of new items and scales, including focus group interviews to investigate the ways in which students interpret the questionnaire and the items;
- Modifying and testing items on a large representative sample of students in Australian universities;
- Statistical analysis of the results and refining of scales;
- Preparing recommendations for the Advisory Committee on potential items and scales.

The project was commissioned in response to growing concerns within the higher education community that important dimensions of the student experience were not tapped by the CEQ. These included the availability and quality of learning resources, in particular, information technology based services and resources, and the extent to which students were engaged in a community of learners. In addition, the consultations with the stakeholders raised the possibility of addressing lifelong learning outcomes they felt were essential for graduates. From these stakeholder meetings a number of similar themes emerged as the basis for the development of items. Contributors suggested a range of areas that might be addressed by an extended CEQ. These can be generally summarised as themes concerning:

- resources and support systems contributing to student learning;

- the quality of the wider student experience;
- independence and autonomy in approaches to learning;
- the perception of the course as an integrated whole;
- levels of intellectual interest, challenge and stimulation;
- the extent to which the course content is relevant and up-to-date;
- 'learning to learn' and the development of skills for lifelong learning.

The final five scales recommended for an extended instrument are:

Student Support Scale. Five items concerned with access to, and satisfaction with, key university facilities and services supporting student learning outcomes;

Learning Resources Scale. Five items primarily focussed on the appropriateness and effectiveness of sources of information and course materials;

Learning Community Scale. Five items on student perceptions of the social experience of learning at university;

Intellectual Motivation Scale. Four items that identify perceptions of the impact of the course in inspiring and enabling individuals, as well as a global item enabling students to evaluate their overall university experience;

Graduate Qualities Scale. Six items tapping qualities typically associated with higher order outcomes, especially attitudes and perspectives related to the relevance of the course for lifelong learning.

The existing CEQ items and scales were retained throughout the pilot testing to allow for comparison with existing time series data. The extended questionnaire, with the existing CEQ included in the normal form, did not appear to distort responses to the CEQ items and scales. The statistical analysis suggests that the additional scales do not have an impact on the existing CEQ scales. The new scales provide valid, reliable and stable estimates of student perceptions of the additional dimensions of the student experience.

The additional 25 items expands the instrument to 50 items altogether. While we are aware that this may affect the response rates in some institutions—an issue that was raised by stakeholders—we believe that some attention to the design of the instrument may counter questionnaire fatigue to some extent. However, the experience of the data gathering in this project suggested the need for some rethinking of the nature and timing of the survey process. This needs to be seen against the potential of the valuable information that can be collected by an expanded instrument to inform the sector and institutions on a wider range of student experiences.

Part I: Development of the instrument

1 Origins and objectives of the project

1.1 The project brief

The Course Experience Questionnaire (CEQ) has been used for the past seven years to survey all graduates from Australian universities in the months soon after their graduation. The CEQ measures aspects of the quality of teaching and learning and the development of generic skills. It comprises 25 items. One item measures overall graduate satisfaction with their course experience, the others can be used to generate five scales*: good teaching, clear goals, appropriate workload, appropriate assessment and generic skills.* The CEQ is considered a valuable instrument for the purpose of improving the quality of teaching in universities and also for informing student choice, managing institutional performance and promoting accountability of the higher education sector.

This project was commissioned in response to concerns within the higher education community that important dimensions of the student experience were not tapped by the CEQ. The brief from DETYA was to identify other aspects of the student experience that an extended instrument might measure, while maintaining the integrity of the existing instrument. The impetus for enhancing the CEQ from the perspective of DETYA is recognition of the diversity of the Australian higher education system and the value attached to the full variety of educational and social experiences of students. The project aimed to develop the CEQ so that it measures broader dimensions of the university experience, while accommodating the diversity of the sector.

The project commenced in July 1999. The main elements were:

- Extensive consultation with the higher education community in preparing the conceptual framework guiding the development of new items and scales;
- Retention of existing items and scales, to permit comparison with existing time series data;
- Trialling of possible new items with current students, including focus groups interviews to investigate the ways in which students interpret the questionnaire;
- Thorough psychometric testing of new items and scales.

1.2 The use and purpose of the CEQ

The Course Experience Questionnaire (CEQ) is included alongside a Graduate Destinations Survey (GDS) as part of the national survey of all university graduates conducted annually by the Graduate Careers Council of Australia (GCCA), The results of the two questionnaires are reported separately. The CEQ provides a national performance indicator of the quality of teaching and is the major source of comparative data on student satisfaction with the overall course experience. This information base is funded by the Department of Education, Training and Youth Affairs (DETYA) and supported by the Australian Vice-Chancellors' Committee (AVCC).

The CEQ asks students who have just completed their undergraduate degree to agree or disagree (on a five point scale) with 25 statements related to their perceptions of the quality of their overall course. The results are reported course by course for every university and have been widely used to support internal quality assurance processes. The questionnaire items have been grouped into four scales concerned with teaching ('good teaching', 'clear goals', 'appropriate assessment', 'appropriate workload'); a scale concerning the acquisition of generic skills for the workforce; and a single item on satisfaction with the quality of the course overall.

While classroom instruction is obviously an important part of the learning environment provided by universities, it is far from the sum of the university experience for students. The CEQ does not account for the social dimension of the student experience and the learning climate that is very much a product of a mix of student attitudes, outlooks and behaviours. Further, the CEQ does not attempt to address 'higher-order' outcomes of the student experience, and gives little direct attention to the area of intellectual stimulation and challenge. Related, and of particular significance to the issue of lifelong learning, is the growing importance of resource-based learning, and the encouragement of independence in learning that must be part of it. A scale from the original items in the CEQ, focussed on the encouragement of independence in learning, was dropped in the GCCA version, apparently in the interests of brevity. There is also a need to measure the provision of resources, guidance and support for students who are encouraged to become increasingly independent in their study. Finally, the CEQ as it currently stands in the GCCA version, does not have the capacity to measure or provide indicators as to the extent to which students take responsibility for their learning.

None of this is intended to suggest that the CEQ is deficient in meeting the objectives set for it. Indeed it has been widely regarded as an essentially robust instrument described by its architect as working well 'in its primary

function of sifting the best from the worst courses within a field of study, and as an indicator that adds to our practical knowledge of what academics must do to ensure that their students achieve excellent learning outcomes.' (Ramsden 1996)

1.3 Assumptions underlying the development of an extended instrument

The initial development phase in this investigation had its origins in a conceptual framework based on extensive research into the impact of university education on student outcomes. This research, much of which comes from the United States, establishes clearly that positive outcomes for students depend on many factors beside classroom instruction—factors associated with the social experiences of students, their interactions with other students and staff, and the nature of the learning climate in the institution.

In their comprehensive review of North American research in the area, Pascarella and Terenzini (1991) conclude that a large part of the impact of the first degree, in both intellectual outcomes and personal development, is determined by the extent and content of a student's interactions with staff and student peers in and out of the classroom. Research conducted by McInnis (1993, 1997) and McInnis and James (1995) has confirmed that these findings are relevant to contemporary Australia. In their studies of the first-year university experience, they established the importance of the social context of learning, including the informal learning opportunities that emerge in unstructured conversations with peers and teachers. It seems that the social dimensions of the university experience are particularly important for young, full-time undergraduates.

A further crucial aspect in measuring the quality of the student experience beyond the CEQ dimensions is the area of intellectual stimulation and challenge. From some preliminary work related to this project we gained the impression that many students judge the quality of teaching on more than the basic proficiency of academics in classroom instruction, yet the dimension of stimulus and challenge is notably absent from the conception of teaching embodied in most evaluative instruments, including the CEQ.

2 Overview of method and recommendations

2.1 Consultations

Wide consultation with the academic community was a critical component of the project. The project began with a survey of universities to determine current use of the CEQ at the institutional level as a preliminary to focus group meetings designed to allow universities to comment on the current utility of the CEQ data and its reporting, and to offer suggestions for possible areas in which new measurement would be valuable.

Meetings were held in four states (Queensland, Victoria, New South Wales and South Australia). All universities were invited to nominate representatives to attend the meetings. Teleconferences were conducted for representatives from Western Australia, the Northern Territory, Tasmania and some regional universities which were unable to send representatives to meetings.

All meetings followed a similar pattern: university representatives were invited to comment on the current uses to which CEQ data were applied in their universities and to propose new areas in which measurement might be beneficial.

The project team also conducted focus group and individual interviews with a small number of students early in the project. The methodology for the on site collection of data also provided considerable opportunities to get first hand comments from students as the trials progressed. Individual academics also provided written comments and suggestions on the direction of the project.

2.2 New dimensions raised by stakeholders and in CEQ responses

Comments on the current CEQ provided useful guidelines for the development of new items and rating scales. Although there was some variability in the extent to which the open-ended section was used, there was strong support for its retention. In some institutions the open-ended responses

currently are systematically analysed and used in the quality assurance processes and guide efforts to improve teaching and learning.

We were fortunate to have made available a sample of these open-ended comments from earlier CEQ surveys provided to us by the University of Ballarat and the University of Melbourne. The Centre for Applied Economic Research at the University of Melbourne made available data from the CEQ for 1996 and 1997. Analysis of responses to the two open ended items at the end of the CEQ, 'What were the best aspects of your course?' and 'What aspects of your course are most in need of improvement?', provided a guide for considering dimensions of student university experience not targeted by the current CEQ. Some response themes related to concepts already included in CEQ. These included the importance of:

- links between academic and external environments through activities such as field trips, practicum programs, guest speakers, workshops, working year, visiting professionals, teaching rounds, industry exposure during research, socialising with academics/professionals, teaching and taking labs, internships, student-employee links;
- universities' ability to present course content and processes as useful, accurate and relevant;
- the importance of motivating students' desire to learn independently;
- students' ability to consider guidance while still being able to structure their own academic programs;
- a social environment that fosters a supportive and encouraging learning community;
- universities' attention to factors in students' lives beyond the academic (religious, cultural, social, economic involvement, need for flexibility);
- general university resources and facilities in sustaining and enhancing learning (eg: effective use of university libraries and information technology).

The focus group meetings with stakeholders (including students) provided a wide range of comments and suggestions for broadening the scope of the instrument. The list that follows illustrates the breadth of issues proposed for consideration for an expanded instrument. Participants suggested the instrument should include items concerned with:

- resources and facilities, including educational technology and their contribution to learning;
- the educational program contributing to graduate qualities of graduates such as intellectual independence and skills of lifelong learning;

- the effectiveness of administration in improving the effectiveness of teaching;
- the relevance of social and community aspects of the university experience;
- specific information about the use of the Internet as a research and/or information sharing/ communication tool;
- the use of various methods to make the learning experience more interesting;
- questions about the cohesiveness of formal course structure and whether it was arranged conceptually and practically to be maximally effective for learning;
- items about assessment to target analytical, creative and social dimensions;
- group work, public presentations and environments open to asking questions and engaging in class discussions;
- investigations of opportunities to access information and advice and make choices about one's own course;
- availability of staff for consultations;
- overall organisation of learning;
- the availability of facilities;
- the broader influence of university life upon students, politically, socially, morally, culturally;
- resources and facilities influential in characterising the form or success of the learning experience;
- a 'learning to learn' and 'learning independence' scale; and
- the acquisition of research skills.

Among the issues raised about the proposed extended instrument, some common themes emerged:

- the inappropriateness of current instruments such as the CEQ to distance education and other specific learning fields and some cohorts, eg: mature age, creative/design students;
- the teacher-centredness of the instrument;
- variations and uncertainties relating to respondent characteristics;
- the problem of time lag and generalisation across many years/subjects; and
- concerns about the –100 to +100 reporting scale.

As a result of the consultations several key areas were identified that were of concern to universities. Records of these were made and added to those identified in the literature. These were then worded into draft items with the intention of matching the response scale used with the current CEQ. The

Likert rating scale is used in the CEQ and was expected to be used in extensions of the instrument. This coding scheme does have some limitations but its familiarity with the instrument meant that it should be retained.

In preparing this report we decided for the sake of brevity not to provide a detailed account of the extensive iterations in the process of defining the final scales.

2.3 Summary of item development and testing

Given the requirement to retain the existing CEQ, the major conceptual developments have been in the form of additional dimensions to the CEQ scales of *good teaching, clear goals, appropriate workload, appropriate assessment, and generic skills.* The primary influence in the development of new scales and items has been the responses of the higher education community listed above. In addition, the work of CSHE over the last two years on measuring the total student experience has been a useful basis for developing new items.

As we commenced arrangements to conduct the item trials and pilot surveys it became clear that many universities were in the process of conducting their own student satisfaction surveys, including the use of the CEQ at all year levels. Some institutions expressed concern that the pilot might have a negative impact on the response rates to their own surveys. Other national and institutional surveys conducted by CSHE were showing lower response rates than anticipated.

To address this challenge of getting sufficient data in a short time-period it was decided to change the proposed methodology and trial new items with face-to-face distribution of draft questionnaires at six universities as an alternative to the mailouts at three institutions originally proposed. The on site approach also had the advantage in the trial phase of providing oral feedback on the meanings attached to the items and their significance from the student perspective. A large number of items were developed for the trials.

Pilot survey

The CEQ items were included in the same form and order in the trial instrument as they appear in the GCCA version. A key criterion for including any new items relating to student attributes was the extent to which universities could be said to have a significant impact on these outcomes.

A consistent concern expressed by contributors to our discussions was the importance of developing items relevant to all students regardless, for example, of institutional type, field of study, or mode of enrolment. The applicability of items for distance education students was a particular issue raised in the initial meetings with university representatives. While it was decided not to develop items specifically targeted at distance students, the project team and advisory group agreed that items should be framed in such a way that all students could meaningfully respond to them. It was also considered important to develop items that measured the extent to which students were engaged with their learning environment and their experience of working with other students.

The project team developed a large set of potential new items, guided by the consultations with university representatives and the background issues raised in the project brief. These items were put through extensive rounds of drafting and redrafting with the advice and assistance of the project's advisory committee. Following statistical analysis of the various pilot versions of items and discussions with the Advisory Group, some 62 items were developed for the field trial. These were broadly grouped under the working headings of:

- Learning community: the extent to which membership of the university community contributed to learning;
- Learning resources: the availability and usefulness of learning materials;
- Curriculum coherence: perceptions of the way the course was organised;
- Intellectual stimulation: focused on the notion the inspirational/motivational dimension that students associate with a first-rate learning experience;
- Choice and flexibility: related to the original CEQ version of the independence scale but incorporating the notion of flexibility in course design and delivery;
- Course impact/relevance: Perceptions of the extent to which the course provided foundations for immediate and long-term growth;
- Graduate qualities: items complementing the existing generic skills scale with qualities associated with lifelong learning;
- Student support and facilities: a series of items concerned with perceptions of the broader student experience of the university.

The trial questionnaire containing 87 items (including the 25 CEQ items) was distributed across 16 universities (face-to-face on campus in 8 universities and mailouts to 8 other universities) during October. The mailout universities were selected with a view to tapping a sample of mature-age, part-time and external students. In total, 3691 students responded to the pilot questionnaire. The instrument administered in the pilot study (and in subsequent trials) incorporated the original CEQ items so that relationships between these and

the new items could be monitored. Rasch analysis provided the principal analytical tool for decision-making, though other analytical approaches were used as appropriate. An explanation of the Rasch method and its appropriateness to this application is included in Part II.

The pilot study led to considerable modification of the items under consideration. Items were discarded or rephrased, and additional items were created. The data from this trial allowed for full analysis of the psychometric properties of the items being considered. The psychometric analysis led to a final set of decisions by the project team and advisory group on the feasibility of particular items and scales for incorporation in an extended CEQ.

2.4 Outline of psychometric approaches

In Part II we discuss in detail the nature of the sample used for field trials and the three methodological approaches used during instrument construction and calibration. The first approach used item response theory, and in particular the Rasch model, to calibrate and evaluate the extended instrument. The second approach used various applications of covariance analysis to further examine the structure and stability of the instrument. Finally, classical test statistics including reliabilities and distractor analyses are considered.

The psychometric analysis was conducted at three levels. The analyses explored whether chosen items were capable of accurate and consistent measurement at the item, scale and instrument contexts. Items and scales were needed that would be sensitive to particular institutional characteristics. In addition they would be required to calibrate in such a way that they would be both consistent with previous implementations of the CEQ and add to the information considered to be of value to institutions. In summary:

1. Item level analysis explored:
    - the dimensionality of sub sets of items in terms of its fit to a modelled variable;
    - item and case separation along the sub-variable;
    - item biases;
    - errors of individual item and case estimates;
    - rating scale performance analysis;
    - rating scale selection (‘distractor’) analysis; and
    - item-total (point-biserial) correlations.

2. At a scale level:
   - covariance between scales;
   - reliabilities of scales;
   - scale scores by sample group; and
   - reliability of scales at key levels of aggregation.
3. Instrument level:
   - inter-item dependencies;
   - factorial dimensionality;
   - reliabilities at key levels of data aggregation; and
   - stability of item estimates and factorial structure across samples.

The analysis included the following steps:

1. definition of the variable—'satisfaction with aspects of the undergraduate experience';
2. identification of the domains or strands within this construct;
3. definition of levels within these constructs;
4. developing items to define levels of satisfaction within domains or strands;
5. drafting items;
6. panelling items (language, content, bias, etc);
7. conflation into scales;
8. pilot of items and scales;
9. adjustment on the basis of pilot studies;
10. trials of scales and background data, analysis of development of norms, and levels of increasing satisfaction.

Development of scales to measure traits such as satisfaction with university experience has been controversial. Interestingly much of the controversy has focussed on methods of analysis rather than the construct being measured. Several studies have attempted to address the construct issue, but more recently the idea of error associated with decisions linked to the CEQ and similar evaluative data has focussed attention on the type of analysis rather than the substance of the analysis. This project has been therefore been aware of both issues and has adopted a multi analytical approach. The core business of the project has been the development of a series of scales, based on student responses to items, similar in content and structure to those embedded in the CEQ, but addressing different aspects of undergraduate life. Hence it has fundamentally relied on item response modelling as the core methodological approach. Other techniques such as confirmatory factor analyses have been used as a cross-check, and multi level modelling has been

used to estimate reliability at various levels of aggregation in an attempt to indicate the accuracy of data at different decision making levels. That is to say if the decisions are to be made, based on the average scores for faculty, or field of study (within an institution), or even at institution level (for government level decisions), then an estimate of the accuracy of the data at each of these levels is required.

The variable has to be defined at student level, since it is at this level that the data is collected. The psychometric procedures have therefore focussed at this level using item response modelling. As yet there are no psychometric techniques of measuring group responses to questionnaire items. No matter which method of analysis is used, it is always the case that the data is collected at the student level. Hence the efforts of this project has been made to ensure that the data is accurate at the primary (student) unit of analysis level and that the aggregation of this data helps to make decisions more accurate at each of the relevant levels of aggregation for decision making.

In order to achieve these goals, the new scales were required to have specific properties. First, they were required to have face validity in that the users must agree that the items and the scales are pertinent and relevant to their institutions and provide useable information. Second, the scales must have adequate reliability at the level at which the data is to be used. This meant that scales should have appropriate levels of reliability in a classical sense and that the error variance at aggregate levels of field of study and institution were within acceptable bounds for decision making. If, for example, resourcing decisions are to be made on the basis of scale scores, reliability at the scale level must be appropriate and this should be estimated at the aggregate level for university or faculty level decisions. Third, the scales and scales must have demonstrated construct validity in that the items in any scale must work together as a cohesive manner to measure a single entity. Without this, assessing reliability at any level is problematic. The first of these requirements was addressed by presenting the items and the suggested scales to the Advisory Committee. The second and third requirements were addressed through a range of empirical analyses.

The Rasch Rating Scale and Partial Credit Models were used to examine the latent properties of scales (Andrich, 1978 and Wright and Masters, 1983). This approach differed from that used in the development of the initial CEQ development. which relied more on exploratory factor analysis. There are compelling reasons for a diversity of empirical approaches in this kind of scale development. Part II of this report provides a detailed discussion of these approaches.

Initial pilot studies of item behaviour conducted at Swinburne University, Deakin University and Ballarat University enabled close involvement with the

students answering the questionnaire. This enabled the field workers to gain insights into the student reaction to the instrument and feedback on issues such as areas they felt were missing, over emphasised, ambiguous or redundant. Site visits were organised through the relevant vice chancellor's office and, where appropriate, student unions. Questionnaires were distributed to students in general and branch libraries, canteens and lecture theatres. Students were provided with pencils and asked to return the form to the field worker who would return in a few moments. This method of approaching students ensured that the responses were obtained from a sufficiently representative cross-section of students by university type and field and enabled a sufficiently large sample to be obtained within the time constraints placed on this portion of the project. Conditions for on site data collection were favourable due to the time of year. On site collection allowed face-to-face interaction with students and to make adjustment to the items as a result of this interaction. The success of the procedure suggested that it was likely to produce higher return rates even for the instrument trials. The sample composition could also be examined daily. It allowed a regular monitoring of sampling requirements and strategy. Because of the large number of items on the pilot instrument, four overlapping forms of the questionnaire were used, anchored by the existing CEQ set of 25 items, which remained constant on each of the questionnaire forms. At each campus a target group of 400 students was selected across a range of disciplines using students from 2nd year or above.

External panelling with the project advisory committee complemented the initial item pilot study. Discussion centred on the opportunity to expose chosen items to a group of representatives and specialists. The members of this group reviewed each item for language, ease of administration, capacity to provide appropriate records and match to perceived needs in accessing information and in providing suitable reports.

Further item level pilot studies were conducted at La Trobe University, The University of Melbourne and Victoria University of Technology with the revised organisation of items. As with the first stage of item pilots, a number of rotated forms were used and a similar procedure adopted as for the earlier pilot studies. The project team scanned, analysed and reviewed the items and the data on a daily basis. This dynamic approach to pilot studies informed changes or modifications to pilot study samples and scale organisation. Rasch model analysis of item performance during pilot studies together the daily analysis of sample composition, enabled exploration of possibilities for extending and reorganising scales further.

As a result of the pilot study analysis items not fitting within the proposed scale structure were omitted. Redundancy, item dependence and misfit to the

Rasch model were used as criteria for deleting items. Some items also needed minor changes to their expression and / or format. Student feedback indicated that there was a need for more emphasis to be placed upon non-academic aspects of university life and a need to focus on the creative skills acquired as well as lifestyle changes attributed to university experience. Following the pilot phase, three rotated forms of the trial questionnaire were developed with different randomisation of new items and the existing CEQ items at the beginning of each form. These instruments were used in the major field trials.

Overall, the project method was intensely iterative. Consultation with the Advisory Committee was built-in at every step to ensure that complementary items and scales would meet the needs of the Australian higher education system and would be perceived as meaningful and relevant by users. At the same time, systematic field trials and rigorous psychometric testing were conducted to ensure confidence in the measurement ability of the items and scales proposed.

2.5 The recommended scales

The five scales and associated items listed below were identified as a result of trials and have been proposed as suitable for inclusion in an extended CEQ questionnaire. Suggested scale abbreviations are given in brackets beside the scale name. Numerical labels used to designate each item throughout the report are given in brackets after each item. The item numbers are consistent with those in the Rasch model analyses presented in Part II.

2.5.1 Student support scale (SSS)

These five items are concerned with access to, and satisfaction with, key university facilities and services supporting student learning outcomes.

1. The library services were readily accessible (76)
2. I was able to access information technology resources when I needed them (77)
3. I was satisfied with the course and careers advice provided (79)
4. Health, welfare and counselling services met my requirements (81)
5. Relevant learning resources were accessible when I needed them (39)

2.5.2 Learning resources scale (LRS)

The items in this scale are primarily focussed on the appropriateness and effectiveness of sources of information and course materials.

1. The library resources were appropriate for my needs (75)
2. Where it was used, the information technology in teaching and learning was effective (78)
3. It was made clear what resources were available to help me learn (38)
4. The study materials were clear and concise (40)
5. Course materials were relevant and up to date (71)

2.5.3 Learning community scale (LCS)

These items concern student perceptions of the social experience of learning at university and indicate their sense of belonging to a community where learning with other people is a priority.

1. I felt part of a group of students and staff committed to learning (29)
2. I was able to explore academic interests with staff and students (31)
3. I learned to explore ideas confidently with other people (34)
4. Students' ideas and suggestions were used during the course (63)
5. I felt I belonged to the university community (30)

2.5.4 Graduate qualities scale (GQS)

The items in this scale are focussed on qualities typically associated with university outcomes, especially attitudes and perspectives related to the relevance of the course for lifelong learning.

1. University stimulated my enthusiasm for further learning (50)
2. The course provided me with a broad overview of my field of knowledge (68)
3. My university experience encouraged me to value perspectives other than my own (51)
4. I learned to apply principles from this course to new situations (54)
5. The course developed my confidence to investigate new ideas (55)
6. I consider what I learned valuable for my future (66)

2.5.5 Intellectual motivation scale (IMS)

The four items in this scale identify perceptions of the impact of the course in inspiring and challenging individuals as well as a global item enabling students to evaluate their overall university experience.

1. I found my studies intellectually stimulating (44)
2. I found the course motivating (49)
3. The course has stimulated my interest in the field of study (46)
4. Overall, my university experience was worthwhile (72)

2.6 Summary of psychometric properties of recommended scales

The scales meet key measurement criteria discussed in detail in Part II. In summary, all items fit scales well and have good measurement properties from Item Response Modelling and Classical Test Theory perspectives. The recommended scales function well in terms of congruence with the current CEQ, and their capacity for discriminating between, and maintaining invariant properties across, different groups of students.

It is important to note that instrument analysis and scoring should proceed at the scale not instrument level. As with the testing of the original CEQ, an initial exploratory factor analysis was conducted on the new items. While this analysis suggested the presence of a single dominant factor, there was evidence of other factors in the data. Furthermore, a Rasch calibration carried out on all items revealed that three items show a lack of relationship with a uni-dimensional satisfaction variable (see Section 4.2). Together these analyses indicate that the individual scales each measure distinct dimensions of the student experience and suggest scoring at the instrument level would be inappropriate.

Variable maps prepared from Rasch calibration of the individual scales (see Section 4.3) reveal a good match between the distribution of student satisfaction measures and the spread of response demand for items. In other words, each of the scales contain a spread of item ‘difficulty’ and are able to differentiate well between students whose responses indicate low levels of agreement and those who express higher levels of agreement. The fit of individual items to the Rasch model for each variable (infit mean square unit) are within an acceptable range. None of the scales contains items that draw random response patterns from students. Mean difficulty and fit values for each item are shown in Table 2.1.

Table 2.1: Mean item difficulty and fit values from the Rasch calibration

Scales	Item number	Logit score	INFIT MS	Scales	Item number	Logit score	INFIT MS
GQS	50	0.290	1.04	LCS	29	–0.062	0.84
	51	–0.152	1.06		30	0.320	1.10
	54	0.172	0.91		31	0.147	0.82
	55	0.247	0.82		34	–0.505	1.08
	66	–0.482	1.18		63	0.085	1.12
	68	–0.082	1.00	LRS	38	0.022	1.01
COS	25	–0.207	0.81		40	0.257	0.80
	26	0.382	1.10		71	–0.277	1.02
	27	–0.417	0.94		75	0.022	1.27
	60	0.012	1.00		78	–0.042	0.95
	61	0.200	1.05	SSS	76	–0.322	1.18
IMS	44	–0.022	0.91		77	–0.160	1.04
	46	0.035	0.95		79	0.275	1.00
	49	0.627	0.88		81	0.170	1.07
	72	–0.660	1.24		39	0.025	0.74

Scale reliabilities and validities were analysed (Section 4.3) and the findings are summarised in Table 2.2. The Item Separation index, ranging from 0.0 to 1.0, indicates the extent to which each item contributes a unique amount to the interpretation of the variable. The indices are close to 1.0 for each of the scales, indicating that the items are well separated along the variable being measured. Similarly, the Student Separation Index provides evidence of the capacity of the scales to discriminate between differing levels of student satisfaction. The five-point Likert scales function well for all items.

Table 2.2: Indices for item and student estimates on scales

Scales	Item separation	Student separation	Cronbach α
SSS	0.95	0.70	0.71
LRS	0.92	0.74	0.76
LCS	0.96	0.78	0.80
GQS	0.96	0.79	0.83
IMS	0.98	0.75	0.83

Additional analytical techniques were employed to provide cross validation of the Rasch analysis (Section 5.1). Covariance modelling of the individual scales was conducted (Section 5.2). The intra-scale item correlations (that is, the correlations between items within single scales) are sufficiently high to indicate the items are forming a coherent group, yet not high enough to suggest they are redundant. Single factor congeneric modelling applied to all the items generated a model that fits, though at a less than satisfactory level.

This outcome is consistent with the findings from exploratory factor analysis and Rasch analysis, adding weight to the conclusion that more than one factor appears to underpin the complete set of items. As a test of the construct validity of the scales, each was modelled using a single factor congeneric factor model. Estimates of the regression and unique error parameters as well as fit indices are presented in Table 2.3.

Table 2.3: GFI, RMSEA and single factor item loadings on scales

Scale	Item	Loading on common variable	RMSEA	GFI
Student	76	0.635	0.085	0.984
Support	77	0.643		
Scale	79	0.462		
	81	0.392		
	39	0.750		
Learning	75	0.530	0.086	0.984
Resources	78	0.620		
Scale	38	0.653		
	40	0.676		
	71	0.612		
Learning	29	0.750	0.038	0.996
Community	31	0.732		
Scale	34	0.592		
	63	0.593		
	30	0.667		
Graduate	50	0.697	0.071	0.983
Qualities	68	0.624		
Scale	51	0.641		
	54	0.675		
	55	0.714		
	66	0.665		
Intellectual	44	0.778	0.000	1.000
Motivation	49	0.766		
Scale	46	0.769		
	72	0.668		

Finally, to explore relations between the new scales themselves as well as between the new scales and the CEQ scales, covariation between the scales was examined. This analysis showed that the new scales independently form constructs. Table 2.4 presents the correlations between the scales.

Table 2.4: New and current scale interscale correlations

	SSS	LRS	LCS	GQS	IMS	GTS	GSS	CGS	AWS
SSS									
LRS	1.00								
LCS	0.68	0.65							
GQS	0.57	0.67	0.68						
IMS	0.58	0.69	0.66	0.98					
GTS	0.54	0.60	0.70	0.60	0.65				
GSS	0.47	0.54	0.61	0.83	0.75	0.58			
CGS	0.58	0.64	0.56	0.55	0.59	0.69	0.51		
AWS	0.23	0.23	0.15	0.19	0.20	0.31	0.07	0.39	
AAS	0.20	0.31	0.24	0.42	0.45	0.43	0.34	0.37	0.37

It is important to note that despite the high correlation between some new and current scales, the salient consideration is that the presence of a strong statistical relationship between two variables does not imply they are measuring the same thing. Thus, the relationship of new to current CEQ scales does not imply redundancy, but, conversely, may even imply congruence.

2.7 Issues for implementation, administration and further development

Managing response rates is a major issue for the implementation of an extended questionnaire. National and institutional quality assurance processes in Australia have prompted major efforts to measure student satisfaction. The distribution of questionnaires to students covering virtually every aspect of their university experience, and especially comments on teaching, has now become so common that the proliferation of surveys from a wide range of stakeholders actually threatens to undermine their usefulness. Survey saturation has led to declining response rates matched by rising student and academic cynicism about their relevance, legitimacy and impact and, as a consequence, producing unreliable and often misleading results.

The scales that were rejected for this instrument on psychometric and conceptual grounds should not be discarded entirely. They were developed in response to issues raised and concerns expressed by stakeholders from diverse contexts across the system. In this period of rapid change in higher education, there is a need to regularly review the scope and approach of instruments designed to measure the quality of student learning experiences.

As Pascarella and Terenzini argue: 'The research questions, designs, and methodologies on which we have relied in the past will not be adequate in the future. We must look to adapt them to the changing conditions and to develop new designs and methods.' (1998 p.163). Most existing questionnaires are rapidly dating as the introduction of advanced educational technology, a strengthening international orientation, an increasingly diverse student population, and a different set of student expectations transform our universities. We recommend continuing development of new scales in response to these changes in the nature of the student experience and the need to assess the quality of the experience from new perspectives.

Part II: Scale development

3 Sampling and field trials

This chapter provides an overview of how data was gathered and prepared for analysis, who the respondents were and provides details of the field trials.

3.1 Data collection methods and outcomes

The population for this project was defined as the undergraduate student population of Australian universities. Two data collection procedures were used during trials: on site administration of the questionnaire and mailing the questionnaire to the student's home address. With the support of the advisory committee, an on site approach was piloted at local universities. On site collection was retained in the trial phase as an effective means of collecting a high return rate with quality student responses. It had these benefits:

- low cost;
- speed of collection made it possible to collect, scan, analyse and report on data on the same day;
- direct contact with students reduced the need for further student focus groups;
- dynamic rather than static sampling was possible which facilitated fine tuning of the sample;
- higher response rates and ability to access most campus based students; and,
- consistency of model-data fit for mailout and on site collections.

Permission was obtained from each of the Vice Chancellor's offices to conduct on site surveying at eight universities. Detailed arrangements were made with the planning or equivalent department at each university and advice was sought on the best day, time and location at which to survey students. Instructions to subjects were similar to those used during earlier pilot collections. Completed surveys were returned by overnight courier to the Assessment Research Centre for scanning and analysis of demographic data. These were reported back to the field team together with instructions for targeted sampling in the following day's collection. This dynamic, purposive sampling approach ensured a representative sample of the general student population from each university and ensured that the sample matched the population characteristics mapped by the Australia Bureau of Statistics (ABS). Further details of this are provided later in this chapter.

Mailout procedures were also used to access those student groups thought to be less represented on campus. These included external/distance, part time and mature age student populations. A random sample of student names and address details drawing from later year undergraduates or early year postgraduates was supplied by participating universities. Student names were either supplied directly and confidentially to the Assessment Research Centre or were kept by the relevant institutions who added postage details onto bundles of envelopes sent by to them by the Assessment Research Centre. Students received an envelope addressed to them that contained a reply paid envelope to Centre for the Study of Higher Education and a survey form. Aside from minor rewording to instructions, identical survey forms were used for on site and mailout collections. One university further assisted the survey by including their own cover letter with mailout. After an appropriate period follow up procedures were initiated with the participating universities. Two universities were sent a second set of survey materials to boost response rates. The following table lists universities that participated in pilot studies and field trials.

Table 3.1: Universities participating in trials and pilot studies

Victoria	Queensland
La Trobe University	James Cook University
The University of Melbourne	Queensland University of Technology
Deakin University	University of Central Queensland
Ballarat University	New South Wales
Swinburne University	Macquarie University
Victoria University of Technology	Australian Catholic University
Tasmania	The University of New South Wales
University of Tasmania	University of Wollongong
South Australia	Western Australia
The Flinders University of South Australia	Murdoch University
The University of Adelaide	Curtin University of Technology
Australian Capital Territory	Edith Cowan University
University of Canberra	

As input from Victorian universities was extensive during the pilot studies of items and scales, representation was sought from universities in other states for trials of the extended CEQ questionnaire forms. Response rates for mailout and on site universities are given in the following figures and tables.

Figure 3.1: Number of questionnaires collected for each university

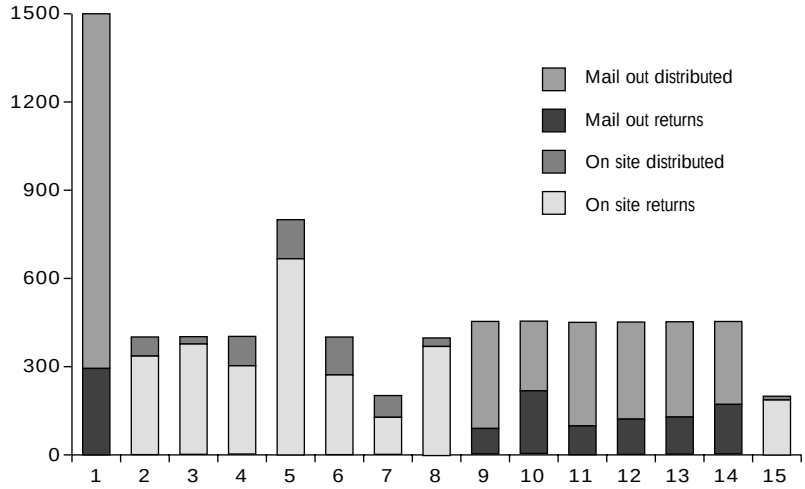
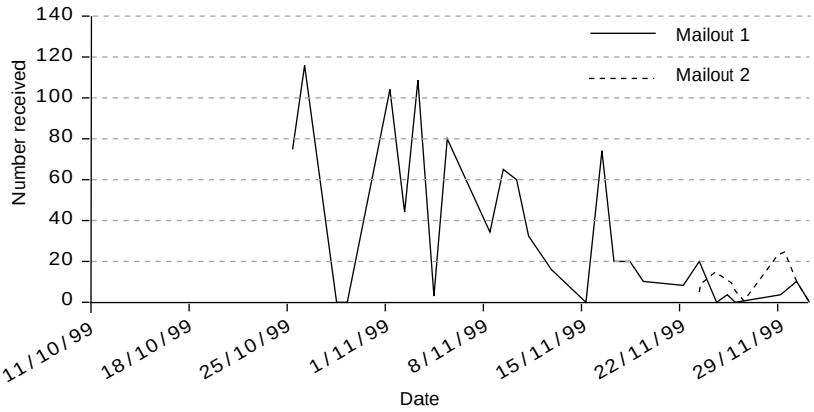


Figure 3.1 shows the considerably higher numbers of questionnaires obtained from the universities where the on site method of data collection was used (70.4 per cent on site compared with 29.6 per cent mailout). Despite the distribution of equivalent numbers of questionnaires to mailout and on site universities, a total of 2632 responses were obtained from on site universities compared with only 1105 from mailout universities. In terms of efficiency, the on site collection process was highly successful. Figure 3.2 shows response rates recorded on a daily basis for both the first and second mailout.

Figure 3.2: Response rate patterns for mailout administration



Despite attempts to boost response numbers by issuing a second mailout, Figure 3.2 suggests a trend of rapidly declining returns and little impact of a second, replacement mailout. The tables below summarise the data collection information. Table 3.2 shows the number of questionnaires received, the number outstanding and the number of responses and questionnaires distributed at each institution.

Table 3.2: Response rates by university

University number	0	1	2	3	5	6	7	8	9
Distributed / Mailed	1500	400	400	400	800	400	200	400	450
Returned	294	335	375	300	667	271	126	371	86
Response rate (%)	19.6	83.7	93.7	75.0	83.3	67.7	63.0	92.7	19.1

University number	10	11	12	13	14	15	Total on site	Total mailout	*Grand total*
Distributed / Mailed	450	450	450	450	450	200	3200	4200	*7400*
Returned	213	98	120	126	168	187	2632	1105	*3737*
Response rate (%)	47.3	21.7	26.6	28.0	37.3	93.5	82.3	26.3	*50.5*

When the data were cleaned and unusable returns were eliminated, a total of 3691 useable questionnaires remained. Checks were undertaken to ascertain whether the patterns of responses were systematically affected by data collection method. For example, item response modelling and confirmatory factor analysis was used to confirm that the instrument performed in a manner consistent with that suggested by other studies of the CEQ (Wilson et al 1996; Johnson, 1997, 1998, 1999). The data also provided evidence of high reliabilities of item estimates and scales, as well as both item and respondent data fit to the modelled variables.

3.2 Data cleaning and editing

December 1, 1999 was set as the final cut off date for instrument return. All 3737 questionnaires received before that date were scanned into the data file. The data were cleaned and edited in preparation for analysis. Although Rasch analysis is robust to missing data, a number of other procedures are not, and it was necessary to use equivalent data sets across all analyses. Thus any cases with missing responses to the current 25 CEQ items or to the 30 new items were removed for analyses across methods. All items and all data were used for Rasch analysis, because of the robustness of the method to missing data. A total of 3691 students responded to the trial questionnaire and 2613 of them provided complete data sets (that is, there was no item with missing answers). The complete data sets were used for comparisons of data

analysis techniques, because procedures such as structural equation modelling are not robust to missing data. The full data, including students who has missed some questionnaire items, were used to establish the properties of the scales. The structure of the data file mitigated against many analyses. The project team was concerned to obtain a sample that was representative of broad fields of study and of institutions, but not of both. Some fields of study (eg medicine) are not available at all universities and this forced some analyses such as Multi Level Modelling to consider empty cells. For this reason many of the analyses were conducted on sub sets of data.

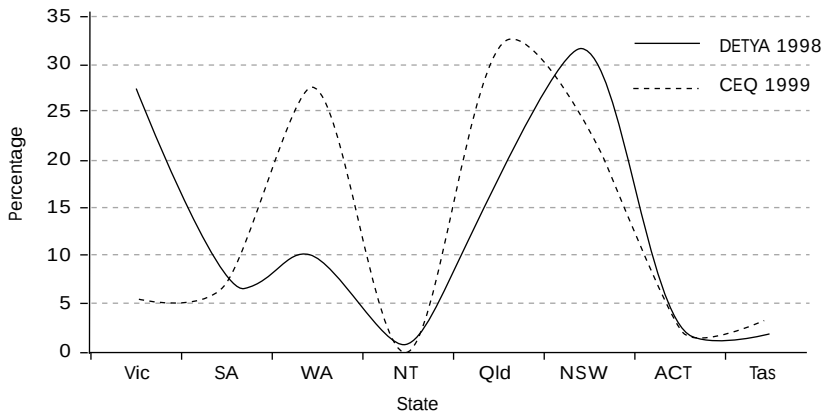
3.3 Sample demographics

The dynamic sampling method used for on site administration of the trial CEQ Extension questionnaire was guided by the need to obtain a sample representative of the target Australian university population. Current data (CEQ 1999) were compared with national figures obtained from DETYA (1998).

Table 3.3: Comparison of sample figures and population estimates

State	DETYA 1998	CEQ 1999
VIC	27.2	5.1
SA	7.4	7.2
WA	9.9	27.4
NT	0.7	0.0
QLD	17.4	31.3
NSW	31.0	23.1
ACT	3.0	2.6
TAS	1.9	3.4

Figure 3.3: Population-sample comparisons for university location



Victorian universities were deliberately underrepresented in the current sample as they were over-represented in extensive earlier piloting. ABS data for fields of study were also used to compare the composition of the current sample. Every field of study is appropriately represented except humanities, which was deliberately curtailed in numbers during trailing also due to overemphasis in earlier pilots.

Table 3.4: Comparison between population and sample figures for BFOS

BFOS	DETYA 1998	CEQ 1999
AGR	1.8	2.4
ARC	2.3	2.4
BUS	24.4	22
EDU	11.2	12.1
ENG	7.6	7.6
HEA	11.4	9.5
HUM	25.0	17.0
LAW	4.6	6.8
SCI	15.7	17.0
VET	0.2	3.3

Figure 3.4: National population and trial sample comparisons for BFOS

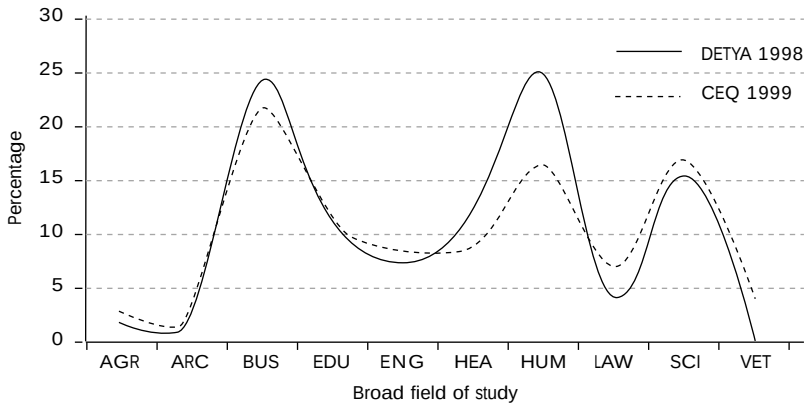
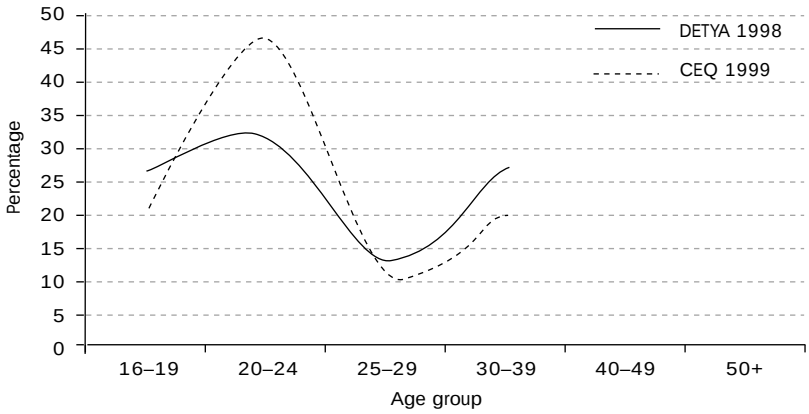


Table 3.5: Population and sample comparisons for age

Age	DETYA 1998	CEQ 1999
16–19	27.1	21.1
20–24	32.4	46.5
25–29	13.2	11.6
30+	27.3	20.8
40–49		
50+		

Figure 3.5: Population-sample comparisons for age



The trial sample is over-representative for the 20–24 age group for university graduates. This reflects an emphasis on campus-based universities and the lack of response from mail out surveys. This is also seen when comparing modes of study, as seen in Table 3.6.

Table 3.6: Sample and national data mode of study figures

Mode	DETYA 1998	CEQ 1999
Full time	59.4	75.7
Part time	27.3	8.7
External or combination	13.3	15.7

Gender is, however, appropriately represented as Table 3.7 shows.

Table 3.7: Comparison of sample and DETYA data for sex

Sex	DETYA 1998	CEQ 1999
Female	54.4	58.9
Male	45.6	41.1

The sample for trials of the instrument closely matched the national population on several key dimensions. The following tables present sample demographics. They relate to the data after cleaning and editing and for cases with complete data used for comparison of analysis methods. The distribution of student responses from each university are shown in Table 3.8.

Table 3.8: Responses from each university

University	Complete data	Percent	Returns	Percent
0	203	7.8	286	7.75
1	204	7.8	335	9.08
2	295	11.3	375	10.16
3	200	7.7	300	8.13
5	485	18.6	667	18.07
6	164	6.3	271	7.34
7	90	3.4	126	3.41
8	229	8.8	371	10.05
9	68	2.6	85	2.30
10	143	5.5	204	5.53
11	67	2.6	95	2.57
12	97	3.7	116	3.14
13	87	3.3	124	3.36
14	117	4.5	149	4.04
15	164	6.3	187	5.07
Total	*2 613*	*70.8*	*3 691*	*100.00*

In Table 3.9, the total number of fields of study exceeds the number of students due to students reporting double majors or multiple fields of study. Thus, of the 3691 respondents there were 379 with double majors and 91 who did not record their major field of study. Of the 2613 students in the complete data set, there were 288 students with double majors and 48 with no field of study recorded.

Table 3.9: Data numbers for BFOS

BFOS	Complete data	Percent	Returns	Percent
Agriculture	82	2.83	95	2.33
Architecture	74	2.55	94	2.31
Business Studies	622	21.44	874	21.47
Education	320	11.03	480	11.79
Engineering	245	8.45	302	7.42
Health	255	8.79	379	9.31
Humanities and Social Sciences	464	15.99	675	16.58
Law	194	6.69	272	6.68
Science	486	16.75	676	16.61
Vet Science	111	3.83	132	3.24
Missing	48	1.65	91	2.24
Total	*2 901*	*100.00*	*4 070*	*100.00*

The study was also concerned with the effect of mode of study of student experience. There were several classifications of study mode. Students could be classified as wholly full time, wholly part time, as on campus or external or some combination of these modes. Table 3.10 presents the breakdown of these data.

Table 3.10: Data frequencies for mode of study

Modes of study	Complete data	Percent	Returns	Percent
Wholly full time	1 966	75.2	2 745	74.37
Wholly part time	235	9.0	322	8.72
Wholly external	120	4.6	185	5.01
A combination of these	269	10.3	388	10.51
Missing	23	0.9	51	1.38
Total	*2 613*	*100.0*	*3 691*	*100.00*

The questionnaire was also presented to students at a range of year levels. Although the CEQ data is collected from students only after graduation, studies at The University of Melbourne have shown that the CEQ is basically stable over year levels. On campus data collection therefore gave an opportunity to gather data from a range of year level and to ascertain whether the year level had an effect on level of satisfaction or on the scales identified in the extensions to the CEQ. This also enabled the project team to examine whether the extensions of the CEQ were as stable as the original CEQ over the year levels. Table 3.11 presents the breakdown of student responses (with complete data) by year level.

Table 3.11: Year level of respondents

Year	Complete data	Percent	Returns	Percent
1	493	18.9	751	20.35
2	766	29.3	1053	28.53
3	758	29.0	1053	28.53
4	362	13.9	490	13.28
5	203	7.8	282	7.64
Missing	31	1.2	62	1.68
Total	*2 613*	*100.0*	*3 691*	*100.00*

Sex and age differences have been shown to be related to satisfaction in a number of studies addressing student satisfaction. Hence sex was included in the demographic data. Tables 3.12 and 3.13 present the gender and age breakdown of the sample providing complete data.

Table 3.12: Gender of subjects

Sex	Complete data	Percent	Returns	Percent
Male	1 088	41.6	1 489	40.34
Female	1 490	57.0	2 137	57.90
Missing	35	1.3	65	1.76
Total	*2 613*	*100.0*	*3 691*	*100.00*

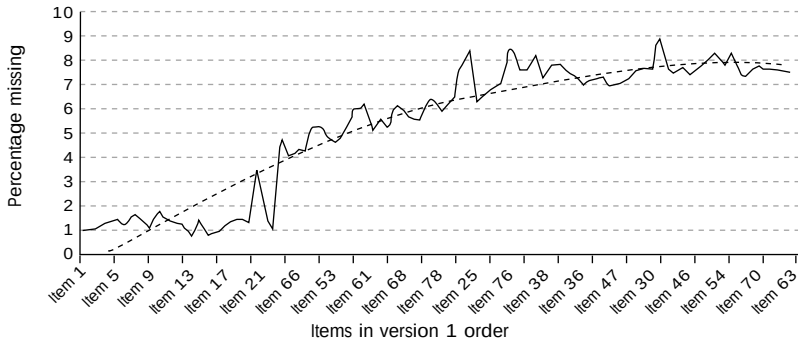
Table 3.13: Respondent age figures

Age	Complete data	Percent	Returns	Percent
16–19	539	20.6	764	20.70
20–24	1 240	47.5	1 681	45.54
25–29	298	11.4	421	11.41
30–39	284	10.9	419	11.35
40–49	199	7.6	301	8.15
50+	21	0.8	32	0.87
Missing	32	1.2	73	1.98
Total	*2 613*	*100.0*	*3 691*	*100*

3.4 Trial instrument characteristics

The psychometric performance of the instrument used during trailing was analysed. The Likert scale used in the initial CEQ was retained in the extension. However only one response set was obtained for each item, rather than two. This is unlike the working version of the CEQ which provides one for major and minor fields of study. Despite the high number of items (25 CEQ items and 62 new items), it was decided that all items would be given to all students. The current CEQ items were presented on one side of the questionnaire in their original order. The rear of the sheet contained a randomised list of all new items being trialled. To counteract length, minimise response set contamination, reduce repetitive fatigue effects on single items, adjust for order effects and allow for later fatigue effect analysis, three different randomisations of the new items were used. Equal numbers of each of the three questionnaire forms were distributed to students. Roughly equal numbers of each questionnaire version were returned (1267 version 1, 1201 version 2 and 1223 version 3). Investigations of respondent behaviour focussed on seeking evidence of response fatigue. This was considered important for a number of reasons. The trial forms of the questionnaire contained more than 80 items and it was considered that such a long instrument would interfere with the reliability of responses. Fatigue effects were examined by observing both the variation in missing responses and fit of the item response patterns to the Rasch model as functions of item order on the questionnaire. Figure 3.6 shows the percentage of missing responses for items in the first randomised form. Figures 3.7 and 3.8 present the examination of fit for a representative set of items.

Figure 3.6: Percentage of missing responses for each item in trial form 1



Patterns for the other forms of the instrument were similar. There was a general increase of the number of missing responses from around 1.0 per cent at the start of the survey to around 8.0 per cent towards the end. A sharp increase is also observable from the 25th item onwards. The 25th item was positioned at the beginning of the second page. This, and the increase in the number of missing responses for items towards the end of the questionnaire, suggests fatigue effects were present. Items on the first page (the current CEQ items) had significantly fewer missing responses than items on the second page (the new items). It was anticipated that giving students 87 items would have such consequences. It was in an effort to counter this, and also to control for order and repetitiveness effects, that three different versions were administered. The intention was to balance the possible fatigue effects across all items by varying the position across items.

The Rasch model is robust to missing data when calibrating items. It allows consideration to be given to the impact of the missing data on the analysis itself. Figure 3.7 presents the fit of items 25 to 28 across trial forms of the questionnaire. The figure illustrates item fit to the model against a scaled down index of item location in the survey. This graph, alongside those involving other items, suggested some, but unimportant, levels of association between item position and fit of the item to the modelled variable. Figure 3.8 shows patterns over all item misfits to the psychometric model.

Figure 3.7: Variations in fit of four items across questionnaire versions

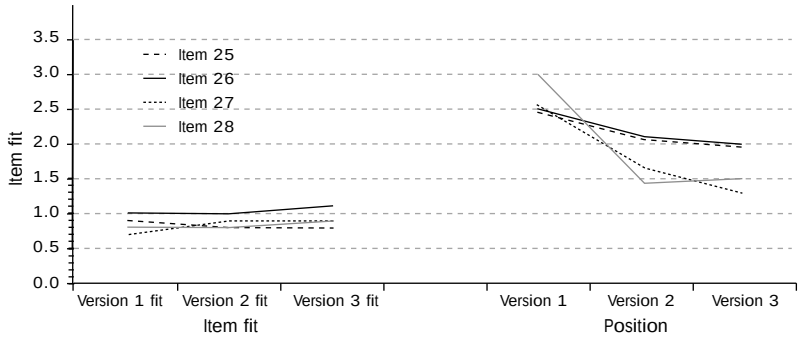
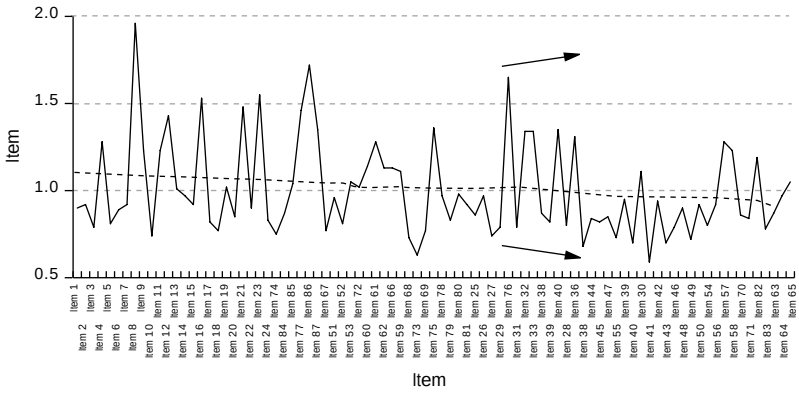


Figure 3.8: The misfit of each item as a function of position in the survey



The vertical scale is a measure of fit to the Rasch model. The horizontal axis represents item numbers given in order of presentation on the questionnaire. The fit statistic is expected to be between +1.30 and +0.77, with an expected value of 1.0. Measures higher than +1.30 would indicate that the response pattern was becoming increasingly random. Fit below +0.77 would indicate that the response pattern was becoming more deterministic perhaps because of non-completion or a fixed response set.

Figure 3.8 indicates that misfit does not increase in value towards the end of the instrument. Trends of item misfit upwards (greater then +1.30) would suggest the increased presence of random (possibly fatigued) response patterns for items later in the questionnaire. Movement downward (less than +0.77) might indicate response set effects. Hence, although there is an increase in the number of missing responses to items presented later in the questionnaire, the psychometric implications of this were minimal.

4 Item response modelling

This chapter presents the Item Response (Rasch) methodology used to model the new scales. Rasch analysis operationalises a scale through estimation of item locations along a latent trait. After presenting the calibrations of all items and the individual scales a brief presentation of results obtained during trialling is given.

4.1 The use of the Rasch model

Wright and Masters (1983) defined the characteristics that measurement should have. They listed four requirements as direction, order, magnitude and replicable units. A questionnaire is a measurement instrument and, like all other measurement instruments, it must be accurate, reliable and have known tolerance or error limits. Estimating these characteristics is called calibration. As Thurstone demanded in 1904, the trait it is measuring should not affect the measurement instrument and the measurement instrument should not affect the trait being measured. Furthermore, the measure obtained of the trait should not be affected by which instrument is used, given that we know the error and accuracy levels of the instrument. Any one of a set of equivalent instruments should give a measure of the trait consistent with measures obtained from with any other equivalent instrument. If a scale is affected by the people who use it or are assessed using it, its validity is threatened. 'Within the range of objects for which the measuring instrument is intended, its function must be independent of the object of measurement' (Thorndike, 1947). This property is known as specific objectivity. For specific objectivity to hold, a measurement must be independent of the measuring instrument and the instrument must function independently of the traits measured. Rasch (1960) first understood the possibilities for objectivity in measurement. He proposed a logistic response model with one item parameter and one person parameter and was the first to apply this type of model to the analysis of test data. The approach requires that we know what it is that we are measuring, that there are units of measurement and that we can link the probability of obtaining a specific answer to a specific item to a persons' ability in the domain we are measuring.

There are several things that are pre-requisite in this. The domain of interest needs to be operationalised. To achieve this, tasks or items are required that can be ordered in terms of the amount of attitude or opinion they demand in order to agree with the substance of the item. This operationalises the

variable. When a scale is argued to measure satisfaction with a course, for instance, the variable (satisfaction) is defined and questions or items that are indicators of the satisfaction are located at different levels on the variable.

The location and interpretation of the levels on the variable represent the contribution of an item response model analysis over other forms of empirical calibration. Rasch analysis assists in constructing a variable (measure) from a dominant dimension (factor) in the data. This dominant dimension may be a hybrid of several constructs, as in the case of the CEQ. In these circumstances, the dominant factor will reflect a composite scale (satisfaction). Lesser dimensions are reported as misfit to the Rasch model. Conducting a factor analysis first using the original observations can lead to misleading results because when observations are non-linear they can, according to Wright (1996), generate illusory factors. Linacre (1998) also argues that exploratory factor analysis can report items clustering at different performance levels as different factors and that there is no way of knowing from factor analysis alone whether each factor is a dimension or a slice of a shared dimension. Factor loadings are correlations of existing data with a latent vector constructed to minimise residuals but the loadings are not on a linear metric. They constrict between –1 and +1 and any plots they may be cast in are coordinates rather than maps of a variable. No approach to factor analysis provides measures of location on the variable and hence prohibit the interpretations of levels of satisfaction.

Rasch analysis constructs linear measures and helps to identify a core construct inside a milieu of co-linearity. This is the message of a 1996 issue of Structural Equation Modeling which contains four articles on the connections between Rasch and Exploratory Factor analysis. Wright (1996, 34), for instance, noted that both misfit to the Rasch model (dimensionality) and the extremities of the unidimensional variable are reported as minor factors by principal component analysis. Further factors are produced by fluctuations in measurement error variance. When a factor cannot be confirmed or established by Rasch analysis, its existence is doubtful. Smith, (1996) used simulation to investigate which technique was better at discovering dimensionality. When the data were dominated by a small number of highly correlated factors, or if one factor dominates, the use of Rasch analysis is recommended. Once a factor has been identified, however, the advice was to separate its items out of the instrument and use Rasch analysis to analyse them further in order to interpret the variable (see Goekoop and Zwinderman, 1994). Chang, (1996, 41–49) also demonstrated that Rasch and Factor analyses produced similar results, but that Rasch results are simpler to interpret, more stable and informative. Factor analysis identifies the relationship to the underlying variable, but not location on it. Rasch analysis, in contrast, provides item and person location on the variable, facilitating the

development of a construct theory and interpretation of levels of satisfaction. A vague factor structure can result in Rasch and Factor analysis suggesting different factors (Green, 1996) . If different variance partitioning, rotation and obliqueness, are used by different analysts, then different factor analyses can produce different factor structures. The nature of a factor solution largely depends on the decisions for the extraction and rotation process, where exploratory factor analysis is concerned. In this project we have not used exploratory fact analysis. Confirmatory analysis however can form an important part of the analysis and strongly supports item response modelling, but still fails to provide interpretative information about levels of satisfaction or measures of location on the variable.

In every attitude scale development procedure the question arises regarding the use of negatively worded items. The Rasch approach does not support this. The discussion above noted that the variable is defined and then items are placed along the variable indicative of the level of satisfaction required to elicit an agreement. Using items as indicators of location obviates the need for considering negated items. Apart from this there is a growing body of evidence that dissuades this practice. Wright (1995) points out that:

> *'NO' is not at all the opposite of 'YES'. What we hate is not the opposite of what we love. Negation releases repression [Freud 1922]. Affirmation acquiesces to custom. Every empathic therapist and father confessor knows that what the patient denies is not the opposite of what they confess… (p.3)*

Interpreting a negation as a positive statement is usually achieved with reverse coding, but according to Wright, this is incorrect. Wright and Masters (1983) display and discuss just such an example. Ebel also wrote about 'true' not measuring the opposite of 'false'. Fifteen years ago and Grosse and Wright (1985) argued against reverse coding approaches to negation in discussing validity and reliability of true-false tests. Negation has not been used in the development of the satisfaction scale. This approach has emphasised dimensionality, accuracy, location on the variable, and interpretability.

The Rasch model is used to predict response patterns to all items by all students. The extent to which this prediction is successful is a measure of the fit of the model to the data. Where the data is shown to misfit, it is a signal that the item data needs to be examined and in most cases, the item is excluded from the scale. The rationale for the omission is based on the assumption that the relationship between satisfaction and the likelihood of a specific response pattern is consistent across items. Where this relationship breaks down, it is assumed that the item is measuring a separate variable. Misfit statistics are standardised mean square differences between observed and expected or predicted values. They have an expected value of 1.0 and

accepted range of values between 0.77 and 1.30. Values below 0.77 are regarded as an indication of an item that is deterministic in nature or that there is dependence on another item, or alternatively, that there is redundancy in the scale. Items with misfit values in this range have a relationship between the student satisfaction and a score category that is too high. Values above 1.30 indicate a low relationship between satisfaction and score. That is, the score assigned does not seem to be related to the overall scale or variable. This could indicate random response, or an item that all or perhaps no one agrees with (the item is too easy or too difficult). The differences between the observed and predicted response patterns are used to calculate the fit statistics. These differences are called the residuals. Residuals are further examined through tests of fit to the model.

4.2 Calibrating all items combined

The trial of the new items and scales was the main purpose of the project. Details of Rasch item analyses are outlined in this chapter. Item response model analysis was used to investigate:

1. fit to the scale variable;
2. item and case dispersion along the sub-variable;
3. compatibility between item and student distributions;
4. reliabilities of individual item and individual case estimates;
5. rating scale performance analysis; and
6. scale dimensionality.

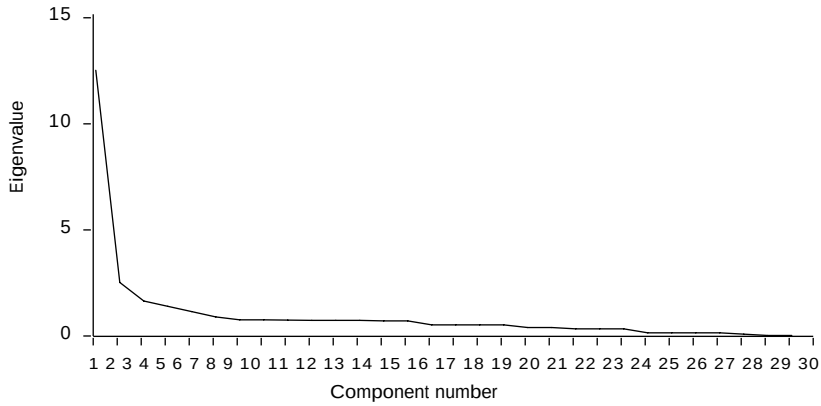
After paneling, pilot studies, item analysis and field trials 30 items were selected for further item and scale level analysis. The 30 items are included below. The number in brackets at the end of each item indicates the item position on the questionnaire and is the reference number used in each of the analyses shown in this report.

1. The library services were readily accessible (76)
2. I was able to access information technology resources when I needed them (77)
3. I was satisfied with the course and careers advice provided (79)
4. Health, welfare and counselling services met my requirements (81)
5. Relevant learning resources were accessible when I needed them (39)
6. The library resources were appropriate for my needs (75)
7. Where it was used, the information technology in teaching and learning was effective (78)

8. It was made clear what resources were available to help me learn (38)
9. The study materials were clear and concise (40)
10. Course materials were relevant and up to date (71)
11. The course was well organised (25)
12. I was given helpful advice when planning my academic program (26)
13. The course content was organised in a systematic way (27,)
14. There was sufficient flexibility in my course to suit my needs (60)
15. I had enough choices of the topics I wanted to study (61)
16. I felt part of a group of students and staff committed to learning (29)
17. I was able to explore academic interests with staff and students (31)
18. I learned to explore ideas confidently with other people (34)
19. Students' ideas and suggestions were used during the course (63)
20. I felt I belonged to the university community (30)
21. University stimulated my enthusiasm for further learning (50)
22. The course provided me with a broad overview of my field of knowledge (68)
23. My university experience encouraged me to value perspectives other than my own (51)
24. I learned to apply principles from this course to new situations (54)
25. The course developed my confidence to investigate new ideas (55)
26. I consider what I learned valuable for my future (66)
27. I found my studies intellectually stimulating (44)
28. I found the course motivating (49)
29. The course has stimulated my interest in the field of study (46)
30. Overall, my university experience was worthwhile (72)

An initial item response analysis was conducted using all 30 items. A principal factor analysis with varimax rotation of the 30 new items suggested the dominance of the initial component extracted (Smith, 1996). It accounted for 34.2% of total variation. This model should account for a considerable proportion of variance in the data and indicates the presence of a single major factor. The scree plot for this analysis is shown in Figure 4.1.

Figure 4.1: Scree plot for principal components analysis of 30 new items



There appears to be one dominant factor, but there are at least one or more weaker factors in the data. It appears that the extension of the CEQ is not unidimensional and that the scales to be presented should be treated separately. However the dominance of the major factor can provide some evidence that the overall satisfaction measure could be used as a guide. Both confirmatory factor analyses and item response modelling were used to follow up this initial exploration of the data.

A Rasch calibration was carried out on all 30 items. The results are presented in Figures 4.2 and 4.3. Figure 4.2 shows the location of the items and students on the variable. The score scale at the left of the figure represents the logit scale. A logit is the natural logarithm of the odds of responding with a specific score category. The score ranges from +3.5 to –3.5, or 7 logits in total. The zero on the scale is arbitrary and represents the average level of the items. Hence all measures on the scale are relative to the mean difficulty (or satisfaction demand) of the items on the scale. The distribution of students on the satisfaction scale is shown to the immediate left of the vertical line. The item score categories are shown on the right of the line. The items are represented by a number in two parts, (X.Y) where X represents the item number, as shown in brackets at the end of each item, and the Y represents the threshold between score categories. The idea of a threshold represents the average satisfaction level required for the score to change from one level to another. For example, 70.1 represents the threshold from a score of 1 to a score of 2, 70.2 represents the threshold from a score on item 70 from a 2 to a 3, 70.3 represents the threshold from a score of 3 to a score of 4 and 70.4 represents the threshold from a score of 4 to a score of 5 on item 70. The figure shows that the distribution of item and score categories matches the distribution of students. The fit of each of the 30 items to the overall scale is

also analysed, and is presented in Figure 4.3. Three items in this figure show a lack of relationship with the variable satisfaction. This finding is consistent with that derived from consideration of the Scree plot in Figure 4.1.

Figure 4.2: Variable map for all 30 items

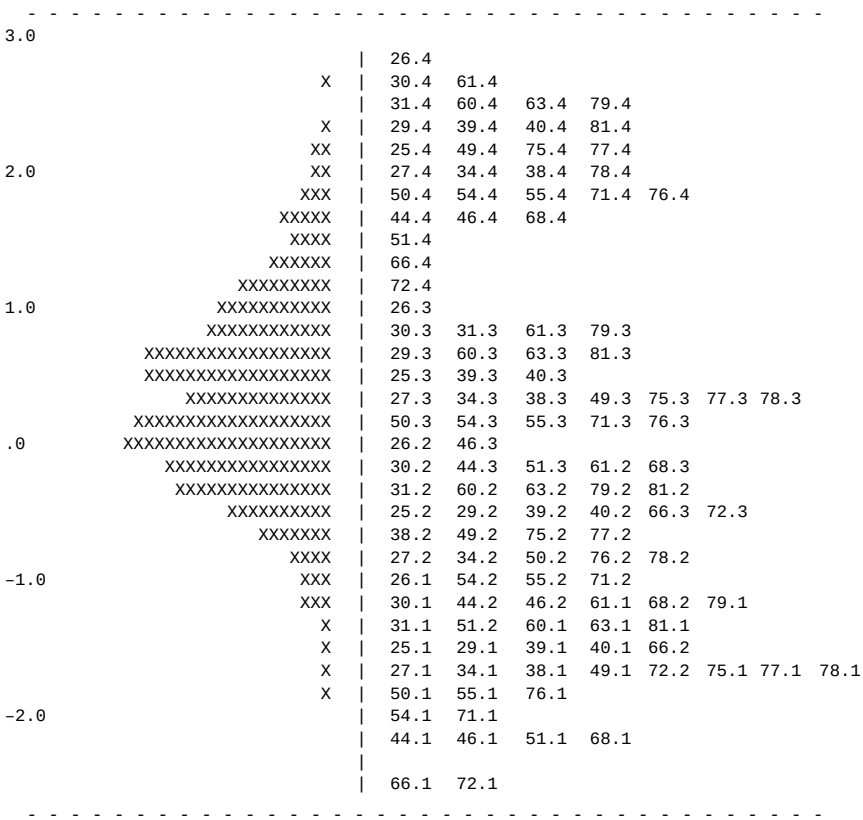
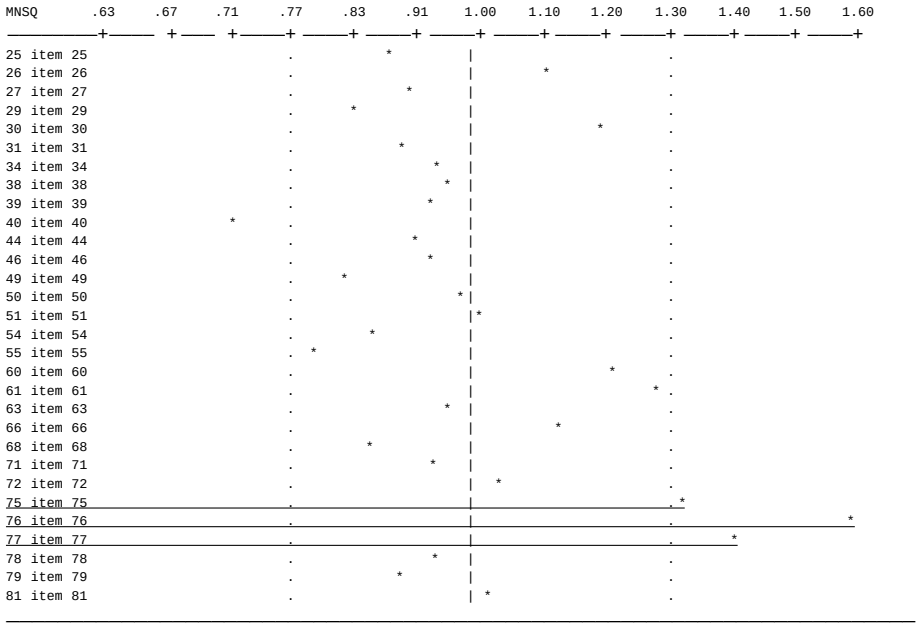


Figure 4.3: Fit of the 30 new items to the overall satisfaction variable



4.3 Calibrating individual scales.

Next, each of the scales was calibrated in turn. The map of modelled variables are shown in Figures 4.4 to 4.15. In these figures, the item numbers are presented on the right of the arrow and students represented as ‘X’s on the left. Each X represents 14 students. As above, the numbers on the right have the form ‘x.y’, where x is the item number and y is the threshold within that particular item matching the position at which response changes from the y–1 to y+1 category of the Likert scale. That is it is the point at which the most likely answer changes from a strongly disagree to disagree (x.1) or from disagree to undecided (x.2) or undecided to agree (x.3) or agree to strongly agree (x.4). The variable increases in the direction of stronger agreement to the items measuring the variable.

The variable maps all indicate a good match between the distribution of student satisfaction measures and the spread of response demand for items. The SSS scale shown in Figures 4.4 and 4.5, for example, has sufficient items to differentiate between students whose responses indicate low levels of student support (76.1, 77.1, 39.1) as well as items to measure students who ‘strongly agree’ with student support items (39.4, 81.4, 71.4). The variable maps also suggest a spread of items along each of the variables, indicating

that there are few positions at which the scales do not match the measures of student attitude.

The infit mean square unit (INFIT MNSQ) in the figures provides a measure of item fit to the Rasch model of the variable against which it is calibrated. The fit of each item is represented as a '∗'. Items with an infit of 1.00 show acceptable fit to the model, items with a fit below 0.77 show patterns of deterministic behaviour in the variable context and items with fit greater than 1.30 patterns of randomness.

Items exhibiting randomness are of greatest concern in the present context. Random response patterns to an item may indicate a lack of association between the item and variable suggesting that the item has poor measurement qualities for that particular scale. None of the new scales contained any such items.

4.3.1 Student Support Scale (SSS)

Figure 4.4: Variable map for Student Support Scale (SSS)

1. The library services were readily accessible (76)
2. I was able to access information technology resources when I needed them (77)
3. I was satisfied with the course and careers advice provided (79)
4. Health, welfare and counselling services met my requirements (81)
5. Relevant learning resources were accessible when I needed them (39)

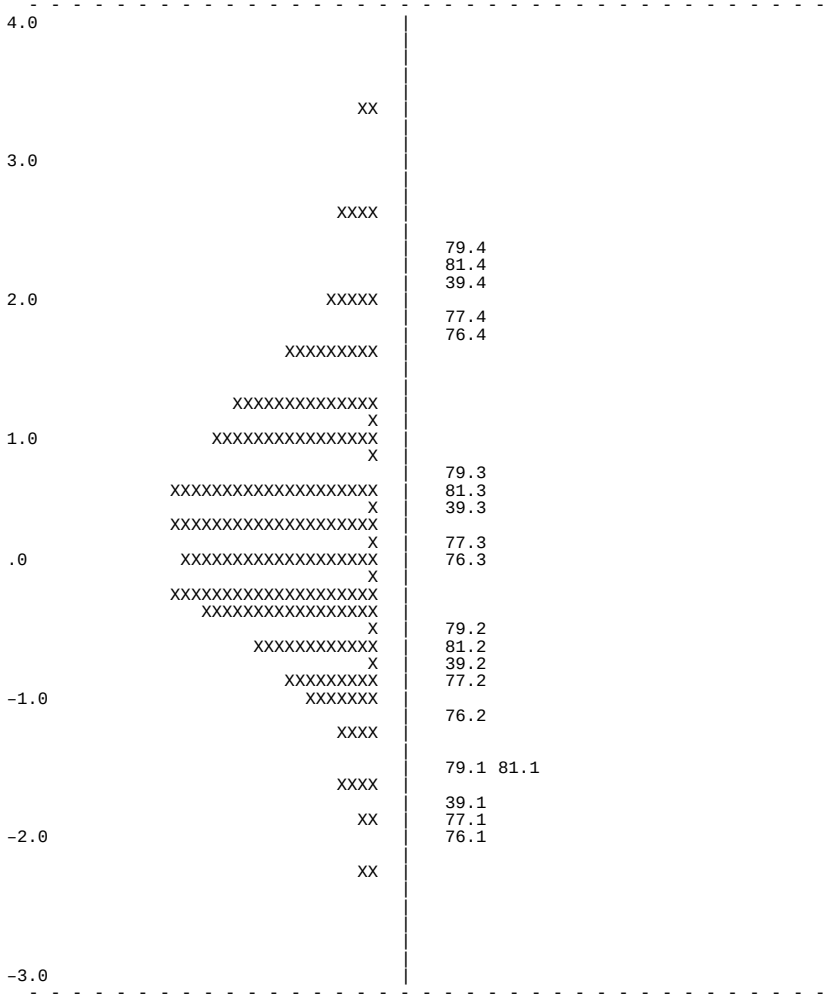
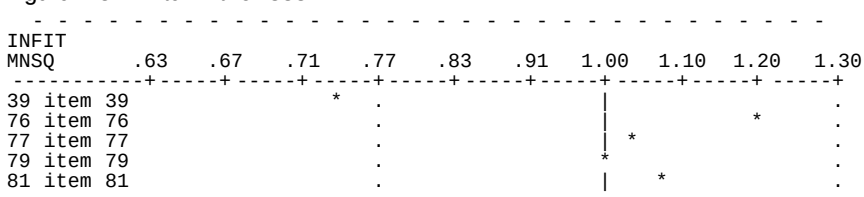


Figure 4.5: Item fit for SSS



4.3.2 Learning Resources Scale (LRS)

Figure 4.6: Variable map for Learning Resources Scale (LRS)

1. The library resources were appropriate for my needs (75)
2. Where it was used, the information technology in teaching and learning was effective (78)
3. It was made clear what resources were available to help me learn (38)
4. The study materials were clear and concise (40)
5. Course materials were relevant and up to date (71)

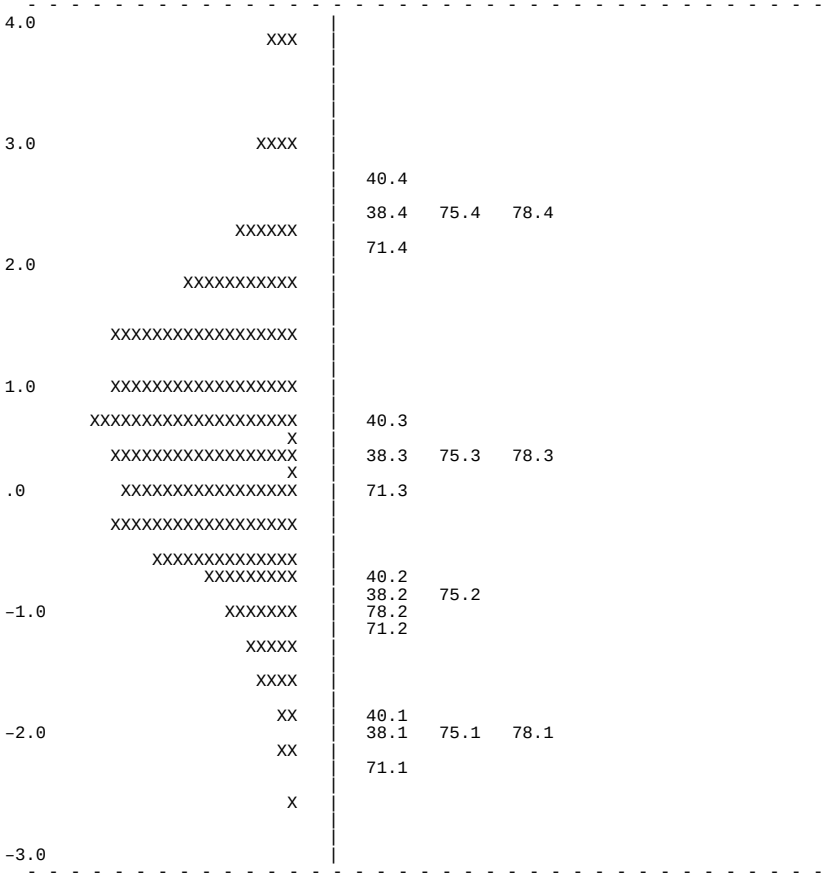
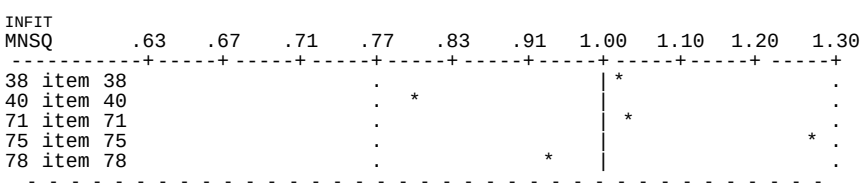


Figure 4.7: Item fit for LRS



4.3.3 Course Organisation Scale (COS)

Figure 4.8: Variable map for Course Organisation Scale (COS)

1. The course was well organised (25)
2. I was given helpful advice when planning my academic program (26)
3. The course content was organised in a systematic way (27,)
4. There was sufficient flexibility in my course to suit my needs (60)
5. I had enough choices of the topics I wanted to study (61)

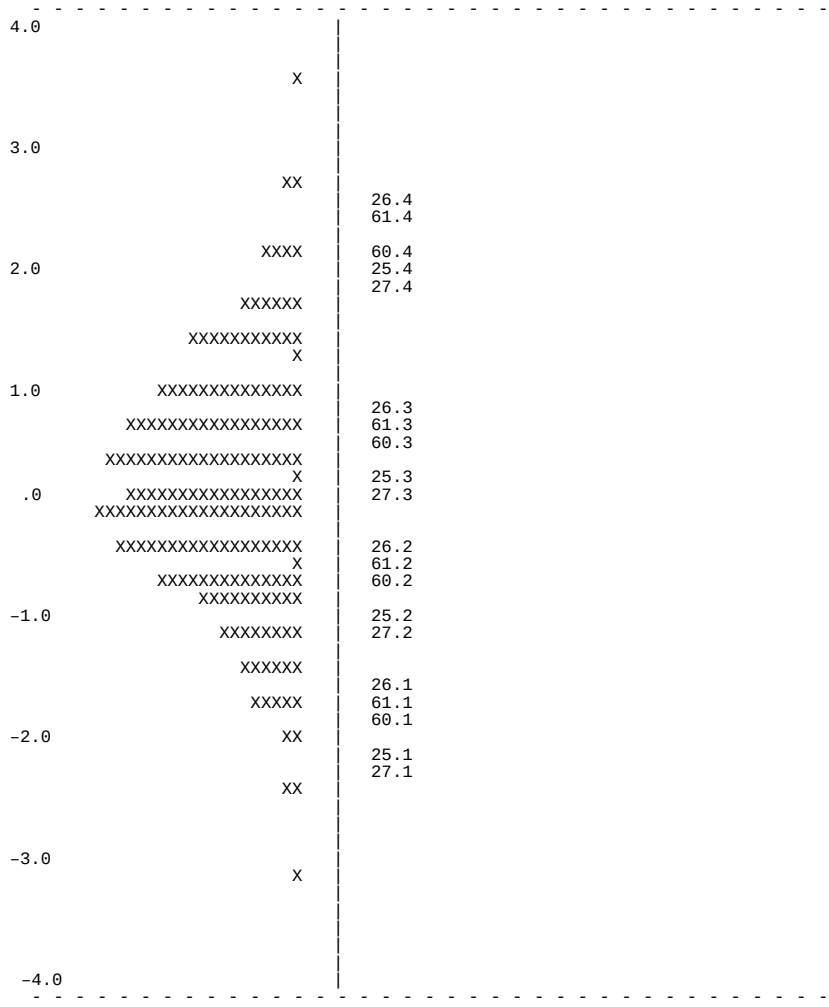
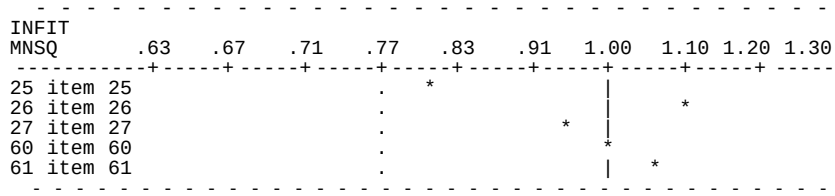


Figure 4.9: Item fit for COS



4.3.4 Learning Community Scale (LCS)

Figure 4.10: Variable map for Learning Community Scale (LCS).

1. I felt part of a group of students and staff committed to learning (29)
2. I was able to explore academic interests with staff and students (31)
3. I learned to explore ideas confidently with other people (34)
4. Students' ideas and suggestions were used during the course (63)
5. I felt I belonged to the university community (30)

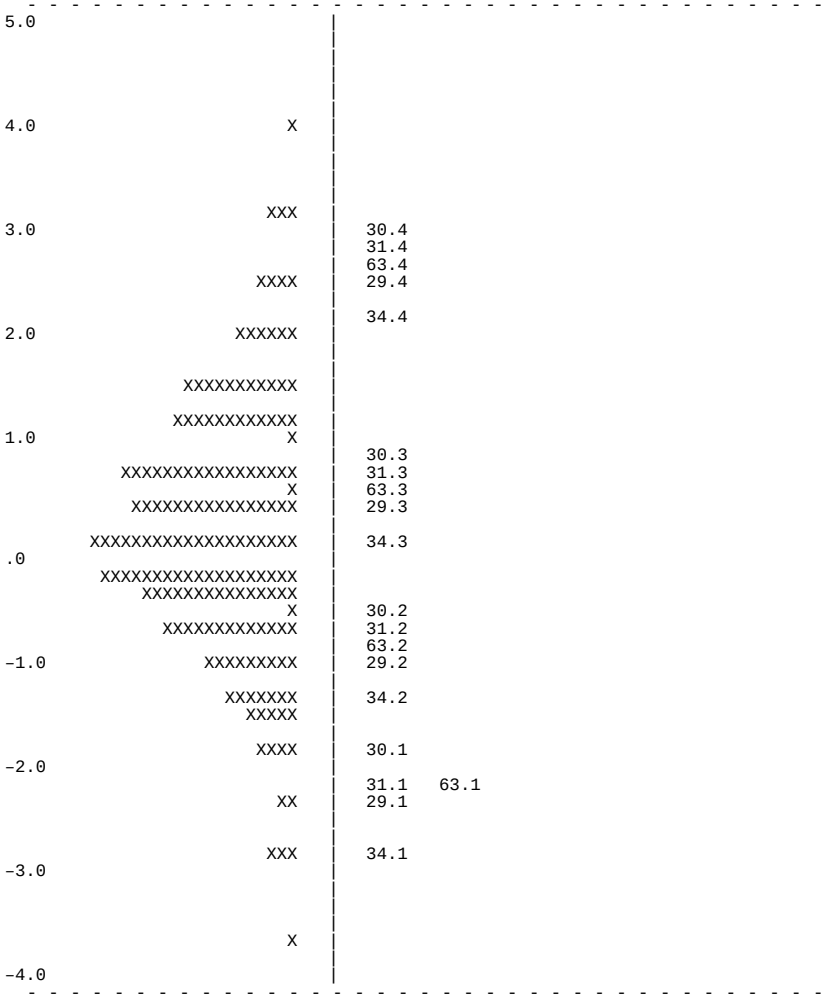
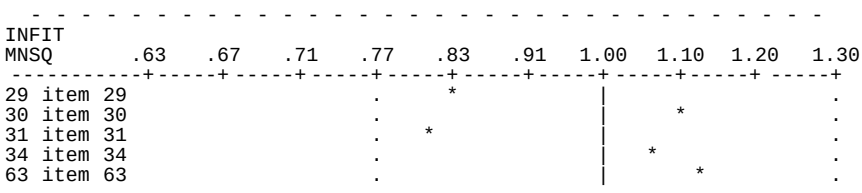


Figure 4.11: Item fit for LCS



4.3.5 Graduate Qualities Scale (GQS)

Figure 4.12: Variable map for Graduate Qualities Scale (GQS)

1. University stimulated my enthusiasm for further learning (50)
2. The course provided me with a broad overview of my field of knowledge (68)
3. My university experience encouraged me to value perspectives other than my own (51)
4. I learned to apply principles from this course to new situations (54)
5. The course developed my confidence to investigate new ideas (55)
6. I consider what I learned valuable for my future (66)

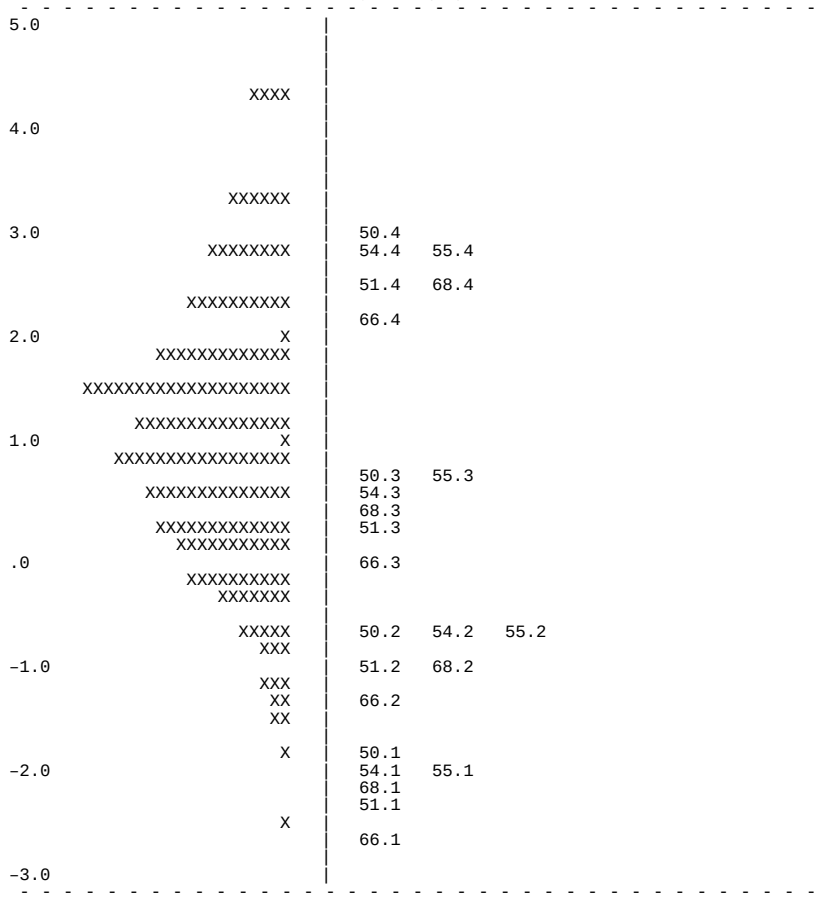
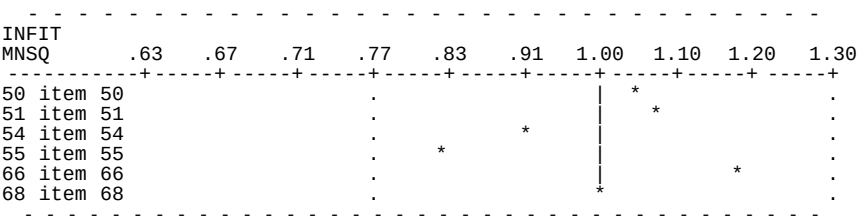


Figure 4.13: Item fit for GQS



4.3.6 Intellectual Motivation Scale (IMS)

Figure 4.14: Variable map for Intellectual Motivation Scale (IMS)

1. I found my studies intellectually stimulating (44)
2. I found the course motivating (49)
3. The course has stimulated my interest in the field of study (46)
4. Overall, my university experience was worthwhile (72)

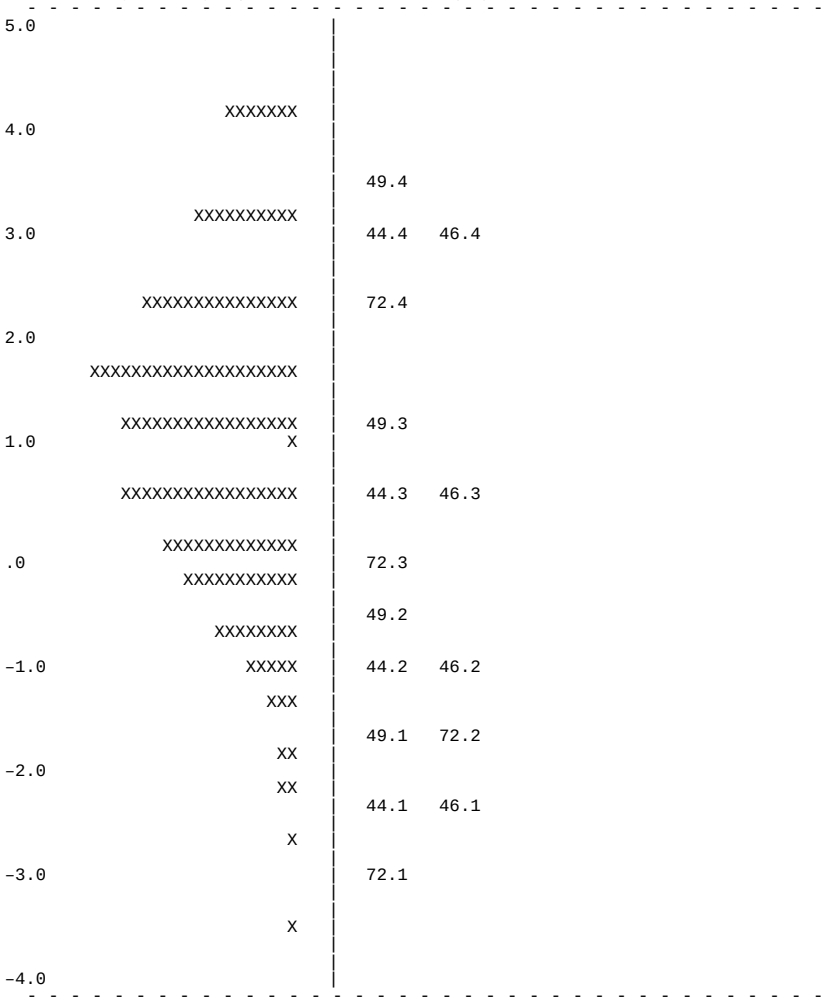
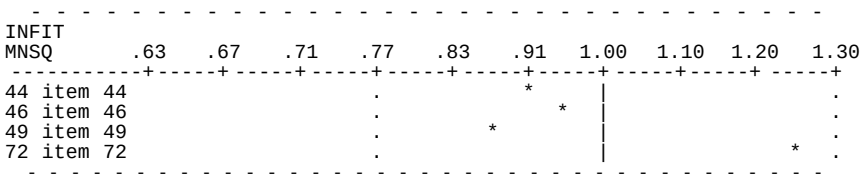


Figure 4.15: Item fit for IMS



Following the item response modelling, scale reliabilities and validities were analysed. Given case and item scores as sufficient statistics, the analysis used item raw scores and student satisfaction raw scores to estimate student satisfaction measures and item demand measures. The items and students were then placed along the same latent variable using the location measure in units called logits. Reliability estimates in a Rasch analysis indicate the extent to which the location measures, given their error estimates, can be separated along the variable. In the case of the item separation indices, the possible range varies from 0.0 to 1.0. A value of 0.0 indicates all items are located at the same position on the variable and that there is complete redundancy in the items' capacities to measure satisfaction. A value of 1.0 indicates that the items (together with their error bands) are completely separated along the variable and each contributes a unique amount to the interpretation of the variable. Wright and Masters (1983) refer to this as the index of construct validity. The estimates are called separation indices. Errors of estimate are available for every item and for every rating category. When combined with a qualitative or content analysis of the items along the variable this provides a way of interpreting the construct being measured. The person (student) separation index is interpreted the same way in the range from 0.0 to 1.0. As the value approaches 1.0, the index provides evidence of the extent to which the scale can discriminate between student satisfaction levels on the variable. This then can be interpreted as an index of criterion validity. Estimates of these statistics for each of the scales are given in Table 4.1.

Table 4.1: Indices for item and student estimates on scales

Scales	Item separation	Student separation	Cronbach α
SSS	0.95	0.70	0.71
LRS	0.92	0.74	0.76
COS	0.97	0.73	0.75
LCS	0.96	0.78	0.80
GQS	0.96	0.79	0.83
IMS	0.98	0.75	0.83

The performance of each item's rating scale was checked, initially through examination of item characteristic curves such as that presented in Figure 14.16.

Figure 4.16: Item 26 characteristic curve

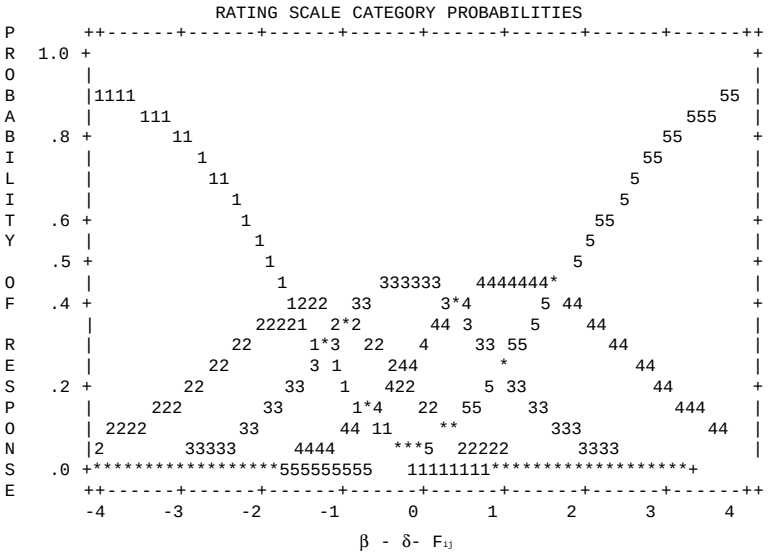


Figure 4.1 shows the relationship between the probability of a response being chosen (from strongly disagree (1) to strongly agree (5) and the difference between the student satisfaction and item demand (β–δ). For a given item, the higher the level of satisfaction the more likely the higher item score category chosen. This chart is typical of all item characteristic curves for the set of items remaining after trials and item analyses. Each response category is, at some stage, the most likely. That is the curve for each category is higher than all other categories. If charts for any particular item had response categories, which were not at some point the most likely response choice, this would suggest that fewer response categories could be used. Likert scales functioned well for all items.

The item characteristics are summarised in Table 4.2 on the following page.

Table 4.2 Summary of Rasch analysis item statistics

Items	β_1	β_2 $\delta_{1,2}$	β_3 $\delta_{2,3}$	β $\delta_{3,4}$	β_5 $\delta_{4,5}$	INFIT MS	SCALE
50 University stimulated my enthusiasm for further learning	–1.05	–0.29 –1.91	0.43 –0.67	1.41 0.75	2.59 2.99	1.04	Student Support Scale (SSS)
51 My university experience encouraged me to value perspectives other than my own	–1.40	–0.52 –2.34	0.20 –1.11	1.23 0.31	2.29 2.53	1.06	
54 I learned to apply principles from this course to new situations	–1.33	–0.48 –2.03	0.36 –0.78	1.41 0.63	2.60 2.87	0.91	
55 The course developed my confidence to investigate new ideas	–1.32	–0.52 –1.97	0.39 –0.70	1.48 0.72	2.72 2.94	0.82	
66 I consider what I learned valuable for my future	–1.17	–0.72 –2.69	0.05 –1.44	1.02 –0.03	2.15 2.23	1.18	
68 The course provided me with a broad overview of my field of knowledge	–1.44	–0.50 –2.28	0.24 –1.04	1.26 0.38	2.39 2.61	1.00	
25 The course was well organised	–1.64	–0.78 –2.09	–0.19 –0.95	0.62 0.24	1.44 1.97	0.81	Learning Resources Scale (LRS)
26 I was given helpful advice when planning my academic program	–1.04	–0.44 –1.50	0.15 –0.35	0.83 0.82	1.48 2.56	1.10	
27 The course content was organised in a systematic way	–1.61	–0.95 –2.31	–0.23 –1.14	0.48 0.03	1.21 1.75	0.94	
60 There was sufficient flexibility in my course to suit my needs	–1.34	–0.63 –1.88	–0.04 –0.72	0.66 0.47	1.43 2.18	1.00	
61 I had enough choices of the topics I wanted to study	–1.18	–0.53 –1.69	0.06 –0.53	0.73 0.65	1.46 2.37	1.05	
44 I found my studies intellectually stimulating	–1.97	–0.94 –2.44	0.14 –1.13	0.15 0.51	2.84 2.97	0.91	Learning Community Scale (LCS)
46 The course has stimulated my interest in the field of study	–1.67	–0.85 –2.38	0.13 –1.06	1.57 0.57	2.89 3.01	0.95	
49 I found the course motivating	–1.66	–0.54 –1.78	0.59 –0.48	2.01 1.17	3.15 3.60	0.88	
72 Overall, my university experience was worthwhile	–1.67	–1.14 –3.06	–0.16 –1.77	1.08 –0.13	2.45 2.32	1.24	
29 I felt part of a group of students and staff committed to learning	–1.90	–0.98 –2.34	–0.09 –0.99	0.84 0.48	2.18 2.60	0.84	Intellectual Motivation Scale (IMS)
30 I felt I belonged to the university community	–1.58	–0.57 –1.94	0.13 –0.62	0.98 0.86	1.95 2.98	1.10	
31 I was able to explore academic interests with staff and students	–1.94	–0.78 –2.13	0.03 –0.78	0.99 0.69	2.31 2.81	0.82	
34 I learned to explore ideas confidently with other people	–2.11	–1.12 –2.78	–0.29 –1.43	0.63 0.04	1.52 2.15	1.08	
63 Students' ideas and suggestions were used during the course	–1.64	–0.73 –2.19	–0.02 –0.83	0.90 0.61	1.87 2.75	1.12	
38 It was made clear what resources were available to help me learn	–1.11	–0.54 –1.97	0.07 –0.82	0.90 0.42	1.92 2.46	1.01	Course Organisation Scale (COS)
40 The study materials were clear and concise	–1.26	–0.47 –1.75	0.17 –0.59	1.14 0.65	2.20 2.72	0.80	
71 Course materials were relevant and up to date	–1.36	–0.66 –2.28	–0.05 –1.12	0.74 0.11	1.75 2.18	1.02	
75 The library resources were appropriate for my needs	–0.92	–0.44 –1.97	0.06 –0.82	0.84 0.42	1.77 2.46	1.27	
78 Where used, the information technology in teaching and learning was effective	–1.34	–0.61 –2.03	0.08 –0.88	0.89 0.34	1.82 2.40	0.95	
76 The library services were readily accessible	–0.97	–0.71 –2.03	–0.16 –1.03	0.49 0.06	1.22 1.71	1.18	Graduate Qualities Scale (GQS)
77 I was able to access information technology resources when I needed them	–1.08	–0.61 –1.88	–0.04 –0.87	0.57 0.24	1.32 1.87	1.04	
79 I was satisfied with the course and careers advice provided	–0.85	–0.34 –1.44	0.14 –0.42	0.83 0.66	1.54 2.30	1.00	
81 Health, welfare and counselling services met my requirements	–0.92	–0.44 –1.53	0.19 –0.54	0.68 0.56	1.34 2.19	1.07	
39 Relevant learning resources were accessible when I needed them.	–1.21	–0.59 –1.69	0.02 –0.68	0.75 0.42	1.64 2.05	0.74	

4.4 Instrument performance

Having outlined the psychometric development of the new items and scales, this section presents a series of figures and tables illustrating the sample characteristics. Figures 4.17 to 4.22 show overall scores across universities, broad fields of study, modes and years of study, age and gender groups. In each of these figures, the vertical axis is in logit unit derivative of Rasch analysis. Logit scores centre on zero as the mean item demand for each scale. A score of more than zero indicates the satisfaction of a group is above the scale mean. A negative score indicates the student group was slightly below the mean value. Caution, however, is recommended when interpreting these charts because the overall instrument is not unidimensional.

Figure 4.17: Overall satisfaction scores across universities

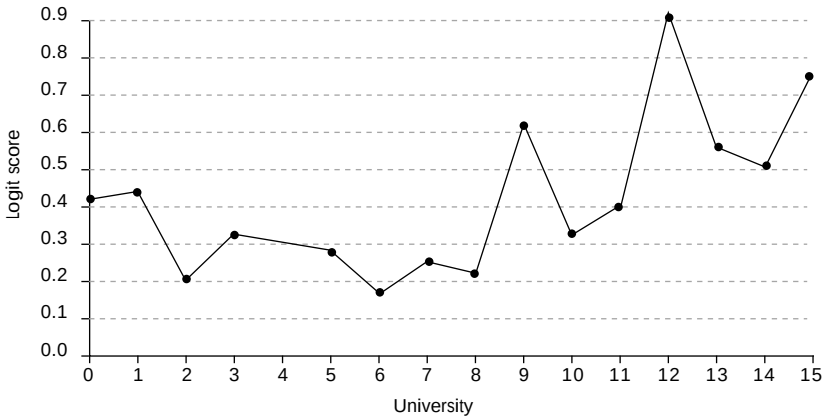


Figure 4.18: Overall satisfaction across broad fields of study

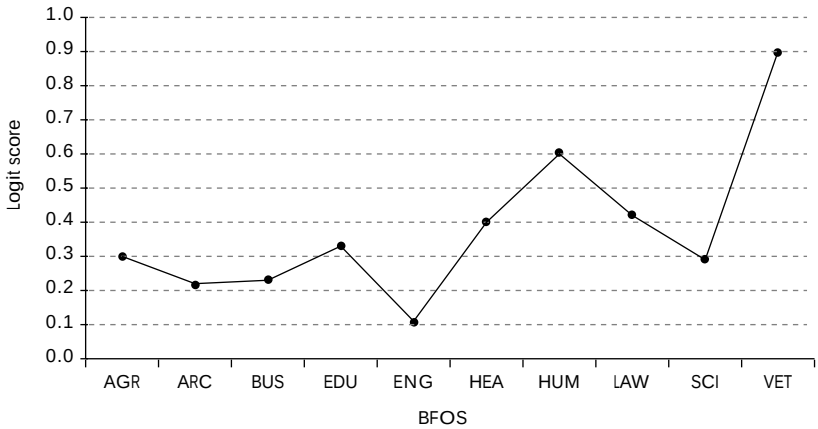


Figure 4.19: Overall satisfaction by mode of study

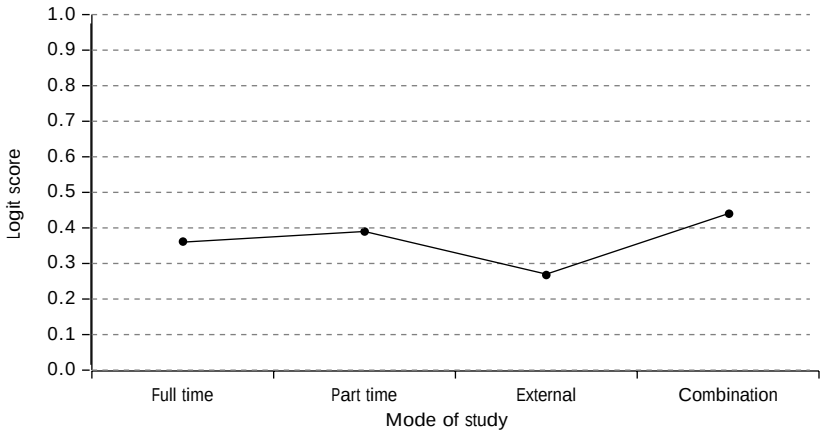


Figure 4.20: Overall satisfaction across years of study with trend line

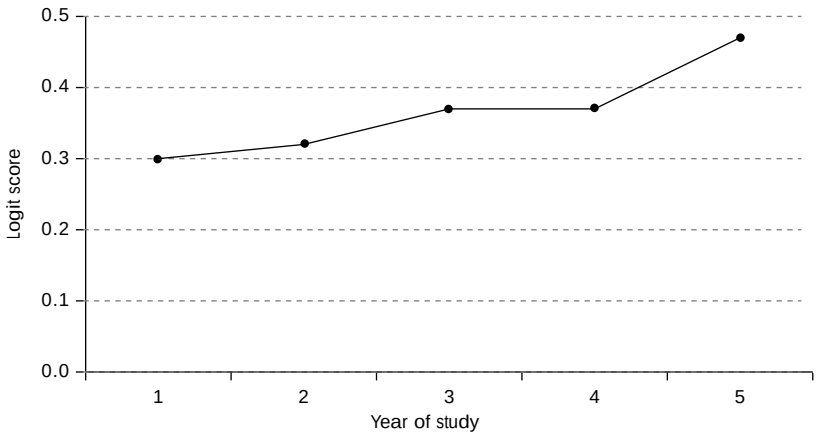


Figure 4.21: Overall satisfaction by age of respondent with trend line

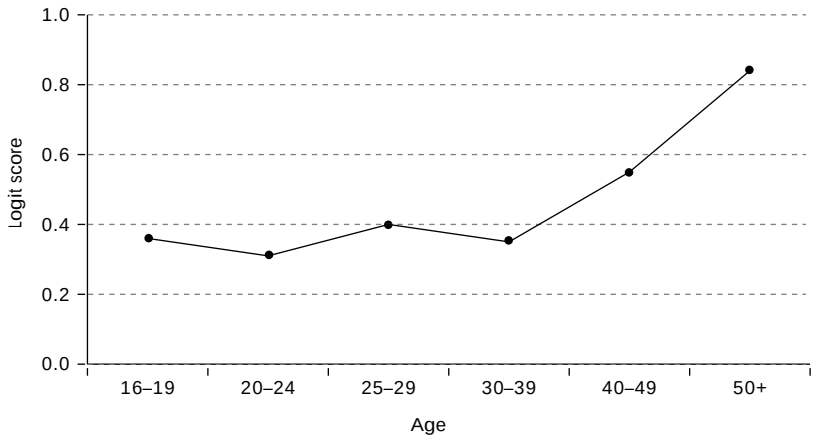
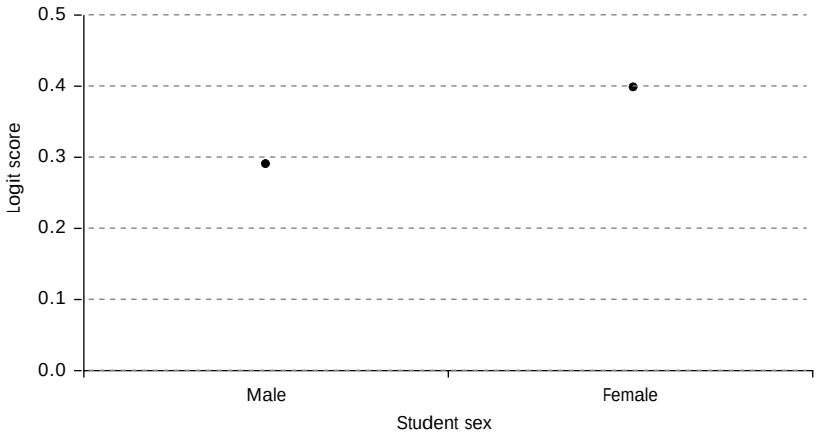


Figure 4.22: Overall satisfaction by sex



Means for individual scales across sample subgroups were calculated. Means are reported in logit units obtained from the item response analysis. Each of the means is shown with their 95 per cent confidence intervals. The intervals are computed from the standard errors. A difference between two means greater than two standard errors implies a possible population difference. However, due to the sampling procedures used in the study, the results need to be interpreted with caution. While it is possible to compare differences between groups on the single scale or perhaps even trends across scales, it is not appropriate to compare differences between or within groups across different scales. The scales measure different things, and these may have naturally different levels of 'satisfaction demand' associated with them. For example it may be easier to get a high level of satisfaction for student support than it is with graduate qualities. Comparative evaluation across scales would require prior standardisation against agreement or satisfaction benchmarks.

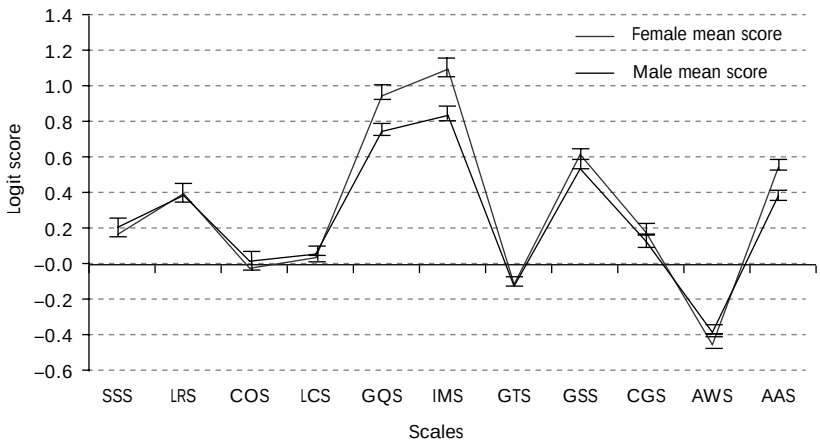
With these observations in mind, Table 4.3 presents mean scores for males and females on each scale. Figure 4.23 plots mean scores for each gender group for each scale.

Table 4.3: Mean and standard errors of scale scores by sex

	SSS		LRS		LCS		GQS		IMS	
Sex	M	F	M	F	M	F	M	F	M	F
Mean	0.21	0.17	0.39	0.40	0.06	0.04	0.75	0.95	0.84	1.10
2SE	0.05	0.05	0.06	0.05	0.06	0.06	0.07	0.06	0.08	0.07

	GTS		GSS		CGS		AWS		AAS	
Sex	M	F	M	F	M	F	M	F	M	F
Mean	–0.12	–0.11	0.54	0.62	0.13	0.18	–0.38	–0.45	0.39	0.55
2SE	0.07	0.06	0.06	0.05	0.07	0.06	0.06	0.05	0.06	0.05

Figure 4.23: Scale scores with 95 per cent confidence bands by sex group



There does not appear to be much variation between gender groups on any of the scales. There appears to be some variation between gender groups on GQS, AAS and IMS scales.

Table 4.4 shows mean differences on each scale by BFOS. Patterns in each scale can be traced over BFOS. For elucidation, a figure showing GTS traced over all ten BFOS is provided beneath the table.

Table 4.4: Mean and standard errors of scale scores by BFOS

		BFOS									
		Ag	Arch	Bus	Edu	Eng	Hlth	Hum	Law	Sci	Vet
GTS	Mean	0.06	0.16	–0.22	0.04	–0.55	–0.06	0.04	–0.05	–0.23	0.47
	2SE	0.21	0.21	0.09	0.13	0.14	0.14	0.12	0.14	0.10	0.20
GSS	Mean	0.64	0.50	0.55	0.69	0.62	0.73	0.61	0.62	0.43	0.74
	2SE	0.21	0.20	0.08	0.11	0.12	0.11	0.10	0.13	0.08	0.18
CGS	Mean	0.19	–0.17	0.23	0.15	0.02	0.19	0.24	0.13	0.10	0.30
	2SE	0.24	0.21	0.09	0.14	0.14	0.14	0.11	0.15	0.10	0.22
AWS	Mean	–0.13	–0.63	–0.35	–0.34	–0.76	–0.45	–0.08	–0.53	–0.49	–1.29
	2SE	0.22	0.25	0.08	0.12	0.14	0.13	0.09	0.14	0.09	0.20
AAS	Mean	0.12	0.56	0.35	0.61	0.35	0.34	0.88	0.81	0.31	0.18
	2SE	0.21	0.23	0.07	0.11	0.12	0.12	0.10	0.13	0.08	0.19
SSS	Mean	0.09	0.09	0.24	0.13	0.10	0.24	0.09	0.04	0.20	0.84
	2SE	0.21	0.16	0.07	0.11	0.10	0.12	0.09	0.13	0.07	0.17
LRS	Mean	0.30	0.05	0.42	0.30	0.23	0.39	0.41	0.32	0.38	1.41
	2SE	0.24	0.20	0.08	0.12	0.12	0.13	0.10	0.14	0.09	0.22
LCS	Mean	0.34	0.19	–0.10	0.21	–0.07	0.12	–0.06	0.05	0.01	0.73
	2SE	0.22	0.25	0.09	0.14	0.14	0.14	0.12	0.16	0.09	0.24
GQS	Mean	0.86	0.76	0.73	0.98	0.60	1.02	1.13	0.91	0.71	1.17
	2SE	0.26	0.28	0.10	0.14	0.14	0.15	0.12	0.17	0.10	0.23
IMS	Mean	1.01	0.66	0.73	0.98	0.63	1.20	1.38	1.04	0.92	1.93
	2SE	0.32	0.30	0.11	0.16	0.18	0.17	0.14	0.20	0.13	0.30

Figure 4.24: GTS scores with 95 per cent confidence bands over all ten BFOS

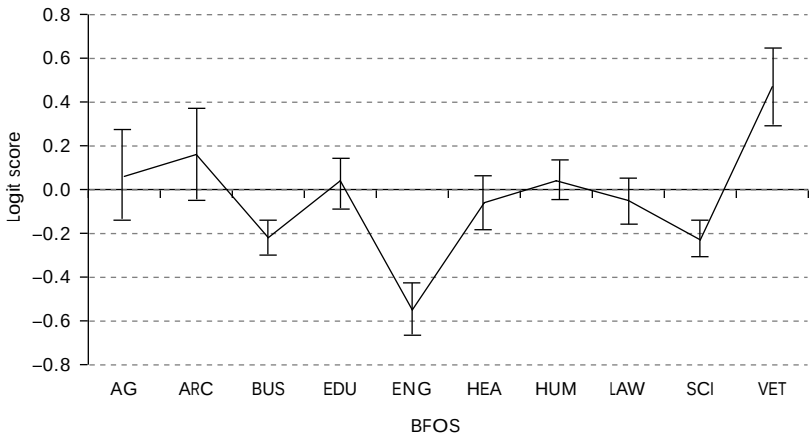


Figure 4.24 illustrates how means can be compared across the ten BFOS. Again, a difference between the 95 per cent confidence bands between any two BFOS suggests a significant difference, if sampling effects are ignored. Most BFOS have mean scores which include the range 0.0 to 0.4. While there are fluctuations between these BFOS, it is possible these are due to chance alone. Veterinary Science has the highest mean score which, aside from Architecture and is significantly higher than any others. Business Studies and Sciences have lower scores than many other BFOS, although Engineering has a significantly lower mean score than others.

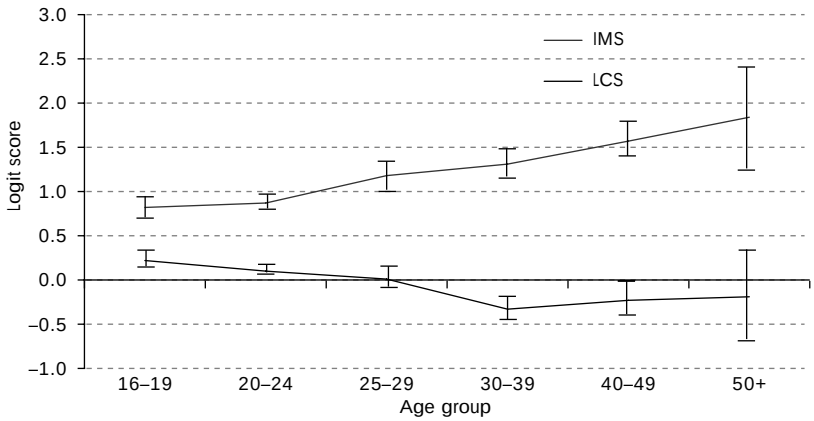
Mean scale scores shown by subject age are presented in Table 4.5.

Table 4.5: Mean and standard errors of scale scores by age of subject

	Age	16–19	20–24	25–29	30–39	40–49	50+
GTS	Mean	–0.12	–0.16	–0.03	–0.08	0.00	0.07
	2SE	0.08	0.06	0.13	0.13	0.18	0.59
GSS	Mean	0.42	0.61	0.60	0.71	0.70	0.72
	2SE	0.07	0.05	0.11	0.12	0.14	0.53
CGS	Mean	0.23	0.10	0.24	0.18	0.16	0.54
	2SE	0.09	0.06	0.14	0.13	0.17	0.47
AWS	Mean	–0.37	–0.50	–0.39	–0.33	–0.23	–0.08
	2SE	0.08	0.06	0.11	0.12	0.13	0.43
AAS	Mean	0.49	0.37	0.52	0.74	0.69	0.94
	2SE	0.07	0.05	0.11	0.11	0.14	0.49
SSS	Mean	0.32	0.18	0.17	0.03	0.07	0.04
	2SE	0.07	0.05	0.10	0.10	0.14	0.32
LRS	Mean	0.52	0.37	0.37	0.23	0.34	0.57
	2SE	0.08	0.06	0.12	0.11	0.15	0.45
LCS	Mean	0.22	0.10	0.01	–0.33	–0.23	–0.19
	2SE	0.09	0.06	0.13	0.14	0.19	0.52
GQS	Mean	0.70	0.78	0.96	1.12	1.41	1.38
	2SE	0.09	0.06	0.15	0.14	0.17	0.57
IMS	Mean	0.82	0.87	1.18	1.31	1.57	1.84
	2SE	0.11	0.08	0.16	0.18	0.20	0.63

The figure below plots LCS and IMS scores over age groups.

Figure 4.25: LCS and IMS scores by age with 95 per cent confidence bands



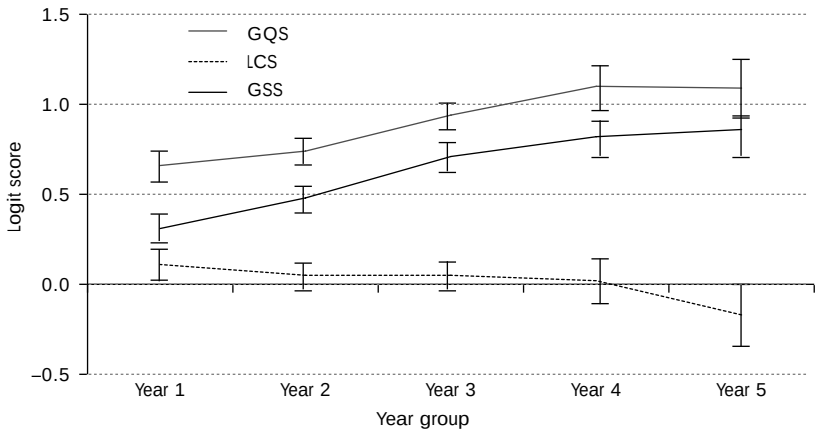
It appears that the extent to which graduates felt intellectually motivated and stimulated by their university experience increases with age. The LCS plot indicates a contrary trend. Older students feel less of a part of a 'university community' (Item 30) which involved them 'working collaboratively with other students' (Item 36). Table 4.6 presents mean scale scores by year of study.

Table 4.6: Mean and standard errors of scale scores by year of study

		1	2	3	4	5
GTS	Mean	–0.10	–0.07	–0.14	–0.11	–0.25
	2SE	0.08	0.08	0.08	0.12	0.16
GSS	Mean	0.31	0.48	0.71	0.82	0.86
	2SE	0.07	0.07	0.07	0.10	0.14
CGS	Mean	0.22	0.15	0.12	0.20	0.02
	2SE	0.09	0.08	0.08	0.12	0.16
AWS	Mean	–0.3	–0.55	–0.45	–0.31	–0.41
	2SE	0.08	0.07	0.07	0.10	0.15
AAS	Mean	0.50	0.38	0.47	0.62	0.56
	2SE	0.08	0.07	0.07	0.10	0.14
SSS	Mean	0.27	0.20	0.13	0.17	0.10
	2SE	0.07	0.06	0.06	0.10	0.13
LRS	Mean	0.43	0.39	0.35	0.41	0.35
	2SE	0.08	0.07	0.07	0.11	0.14
LCS	Mean	0.11	0.05	0.05	0.02	–0.17
	2SE	0.09	0.08	0.08	0.13	0.18
GQS	Mean	0.66	0.74	0.94	1.10	1.09
	2SE	0.09	0.08	0.09	0.12	0.17
IMS	Mean	0.77	0.94	1.04	1.18	1.21
	2SE	0.11	0.10	0.10	0.14	0.19

The mean values for GSS, LCS and GQS are plotted below.

Figure 4.26: GSS, LCS and GQS mean scores with 95 per cent confidence bands



Students' impressions of generic skills obtained through university experience rises with the year of study. Mean scores on GQS also show statistically significant increases over year levels indicating a change in students' attitudes towards their courses generally over years. As students' expectations and application levels change, what appears impractical and irrelevant in earlier years may be emerging as more helpfully related to vocational and extra-curricular activities as students learn that they can 'apply principles from their course to new situations' (Item 54), feel 'stimulated by their course to engage with further learning' (Item 50) and consider 'what they learned as valuable for their future' (Item 66). The figure above also shows that LCS decreases over time and this may reflect the increasingly individual nature of academic involvements that follow higher education academic progressions.

For congruence with current practice, an example report of IMS is given in the format used in annual CEQ reports. This format involves using percentages, counts, means and standard deviations (Johnson, 1997, 1998). Key elements of this reporting process are:

- reporting the percentage agreement and disagreement for each of the five response categories in each item;
- combining the lower two 'disagreement' response categories, the upper two 'agreement categories' into two groups to form, in combination with the middle category, three categories across which results can be aggregated and reported; and
- transforming the five response categories 1 to 5 to (from left to right) –100, –50, 0, 50 and 100, and multiplying by the number of responses to each item to calculate the mean and standard deviation for that item on the CEQ reporting scale of –100 to +100.

Given the match with current methodology, the rationale for this reporting format does not require elaboration. Results are shown in Table 4.7.

Table 4.7: IMS data reported in GCCA format

Item							Mean	SD	N
Response category		1	2	3	4	5			
Rescaled response category		–100	–50	0	50	100			
44	Count	102	296	910	1412	733	34.43	49.69	3453
	%	2.95	8.57	26.35	40.89	21.23			
	% combined	11.53		26.35	62.12				
46	Count	137	303	876	1435	701	32.73	51.13	3452
	%	3.97	8.78	25.38	41.57	20.31			
	% combined	12.75		25.38	61.88				
49	Count	171	478	1161	1202	447	18.44	51.58	3459
	%	4.94	13.82	33.56	34.75	12.92			
	% combined	18.76		33.56	47.67				
72	Count	97	162	686	1378	1139	47.66	49.1	3462
	%	2.80	4.68	19.82	39.80	32.90			
	% combined	7.48		19.82	72.70				

5 Validity checks

Structural equation modelling enables a second approach to the structure of the scales to be explored. Using this alternative methodology provided a cross validation of the results of the item response analysis. It relates to the correlation between the item and a latent variable as the interpretive device. In the item response analysis the location of the item on the variable was the basic interpretive mechanism. Given that both approaches reduce the data set and provide an interpretive mechanism, a cross check of one against the other assists in supporting the construct interpretation of the item set.

5.1 Covariance modelling of all items combined

A single factor congeneric model is a confirmatory approach that tests an hypothesis that a single underlying latent construct is causing the response pattern to each of the items. It further assumes that the variance in the response pattern can be explained by a measurement model derived from the inter-item correlation and covariance. In the Rasch model, the hypothesis examined was that the likelihood of a person response to an item is an outcome of the level of person satisfaction level and the satisfaction demand of an item. The hypothesis is tested using the odds or probability of a response pattern given the predicted measures of both item and person, compared to the observed response patterns. The two approaches are quite different in assumptions and in approach, but if the items do address a specific measurable construct, then the two approaches should provide collaborative evidence of that construct. They both assess the dimensionality of the data and test the fit of the data to a unidimensional model. That is, they both test the evidence of the construct that the item set purports to measure. In the structural equation model, confirmatory factor analysis, each item was forced to load either on a common single factor or on its own unique error term. Low correlations with the common factor (low loadings) and high correlations with the error term combined with poor measures on the Goodness of Fit Index (GFI) and the Root Mean Square Error of Approximation (RMSEA) fit indices indicate lack of fit to the unidimensional model. Table 5.1 below presents estimates for the model with each item regressed on a single common factor.

Table 5.1: Single factor congeneric model applied to all items

Item	Standardised regression weights	Unique error variance
25	0.60	0.67
26	0.49	0.92
27	0.55	0.66
29	0.62	0.65
30	0.49	0.95
31	0.55	0.70
34	0.52	0.67
38	0.57	0.74
39	0.55	0.78
40	0.60	0.56
44	0.69	0.51
46	0.70	0.52
49	0.69	0.55
50	0.65	0.65
51	0.55	0.62
54	0.61	0.54
55	0.63	0.52
60	0.47	0.97
61	0.42	1.06
63	0.52	0.74
66	0.63	0.58
68	0.62	0.51
71	0.57	0.65
72	0.68	0.50
75	0.42	1.01
76	0.41	1.15
77	0.42	1.06
78	0.53	0.71
79	0.58	0.71
81	0.43	0.79

The GFI and RMSEA figures for this model were 0.738 and 0.090 indicating a model that, although it fits, does so at a less than satisfactory level. This is consistent with the interpretations arising from the Rasch and exploratory factor analyses presented in Figures 4.1 and 4.3 above. More than a single factor appears to underpin the complete set of items.

5.2 Covariance modelling of individual scales

Modelling of the relationships within scales was begun by considering intra-scale correlations between items. These are shown in Table 5.2.

Table 5.2: Intra-scale item correlation matrices

SSS	76	77	79	81	39
76	1.0000				
77	0.4438	1.0000			
79	0.2502	0.2689	1.0000		
81	0.2113	0.2277	0.3116	1.0000	
39	0.4789	0.4771	0.3567	0.2886	1.0000

LRS	38	40	71	75	78
38	1.0000				
40	0.4708	1.0000			
71	0.3835	0.4520	1.0000		
75	0.3357	0.3042	0.3073	1.0000	
78	0.3913	0.3875	0.3632	0.4233	1.0000

COS	25	26	27	60	61
25	1.0000				
26	0.4131	1.0000			
27	0.5800	0.3333	1.0000		
60	0.3572	0.3050	0.2794	1.0000	
61	0.3348	0.2859	0.2526	0.5769	1.0000

LCS	29	30	31	34	53
29	1.0000				
30	0.5120	1.0000			
31	0.5300	0.4972	1.0000		
34	0.4619	0.3815	0.4282	1.0000	
53	0.4429	0.3694	0.4615	0.3424	1.0000

GQS	50	51	54	55	66	68
50	1.0000					
51	0.4987	1.0000				
54	0.4507	0.4386	1.0000			
55	0.5310	0.4474	0.4830	1.0000		
66	0.4347	0.3902	0.4551	0.4689	1.0000	
68	0.3926	0.3827	0.4351	0.4200	0.4957	1.0000

IMS	44	46	49	72
44	1.0000			
46	0.5974	1.0000		
49	0.5990	0.5882	1.0000	
72	0.5178	0.5183	0.5096	1.0000

These figures show that items are not highly correlated within each scale. In the student support scale SSS, for instance, the range of intra-item correlations varied from 0.479 to 0.211. For the LRS scale the range was 0.471 to 0.304. In COS the range was from 0.577 to 0.253. For the LCS scale it was 0.530 to 0.342. For the GQS scale the range was 0.531 to 0.392, and for the IMS scale it was 0.599 to 0.510. In most scales the range of intra-item correlations is sufficiently high to indicate that the items are forming a coherent group, but not high enough to indicate that they are redundant. In other words, each item appears to be contributing unique information in a coherent item scale set.

To further test the construct validity of the scales, each of the scales was modelled using a single factor congeneric factor model. Estimates of the regression and unique error parameters as well as fit indices are presented in Table 5.3 for each of the factor models.

Table 5.3: GFI, RMSEA and single factor item loadings on scales

Scale	Item	Loading on common variable	RMSEA	GFI
SSS	76	0.635	0.085	0.984
	77	0.643		
	79	0.462		
	81	0.392		
	39	0.750		
LRS	75	0.530	0.086	0.984
	78	0.620		
	38	0.653		
	40	0.676		
	71	0.612		
COS	25	0.757	0.222	0.906
	26	0.536		
	27	0.653		
	60	0.558		
	61	0.531		
LCS	29	0.750	0.038	0.996
	31	0.732		
	34	0.592		
	63	0.593		
	30	0.667		
GQS	50	0.697	0.071	0.983
	68	0.624		
	51	0.641		
	54	0.675		
	55	0.714		
	66	0.665		
IMS	44	0.778	0.000	1.000
	49	0.766		
	46	0.769		
	72	0.668		

Table 5.3 suggests good fit for each scale except for COS. A statistical approach was made to improving the fit of this scale. Modification indices were consulted and used to suggest the addition of parameters which would improve model fit. In the single factor context of COS, the indices suggested a large improvement in fit if the residual terms of items 60 and 61 were covaried. Both the fit indices approached acceptable values when this was done and the modified model re-estimated. The RMSEA decreased from 0.222 to 0.064 and the GFI increased from 0.906 to 0.993. Covarying error terms in this way can be thought of as clarifying the relationship of the variables to the common factor by partitioning out the variation they do not share with the common factor (Byrne, Shavelson and Muthén, 1989; Byrne, 1994). The approach is defensible as long as the measurement parameters do not undergo large scale change. The relationship modelled between the covaried terms has substantive rationale and the modified model cross validates on new data (Pentz and Chou, 1994). Clearly, however, it also indicates the items measure a secondary construct in addition to organisation. An inspection of the items suggests they may be more about the respondents themselves rather than their courses. Despite fit to the Rasch model, confirmatory factor modelling of this scale suggests that while the items can be modelled to be statistically homogeneous, substantively and thus practically the scale may lack coherence. The robustness of the Rasch analysis might not be sufficient to convince all analysis of the unidimensionality of the COS scale.

5.3 Structural models

To consider relations between the new scales, the data for each of the scales was allowed to co-vary in a six-factor model. The model tested the hypothesis that the data from the item set formed a construct for each scale and these were in turn correlated. This had an acceptable fit with GFI=0.895 and RMSEA=0.061. The hypothesis that the scales were separate entities was supported. Hence each of the scales was then examined using a single factor model. Loadings are presented in the following table for each of the scales. This had an acceptable fit with GFI=0.895 and RMSEA=0.061. The model estimates and interscale correlations are given in tables 5.4 and 5.5. It is clear that the scales need to be treated as separate entities.

Table 5.4: Parameter estimates from structural model covarying new scales

Scale	Item	Standardised regression estimate	Variance estimate
GQS	50	0.72	0.55
	68	0.65	0.48
	51	0.61	0.56
	54	0.66	0.49
	55	0.70	0.44
	66	0.68	0.52
IMS	44	0.76	0.41
	49	0.75	0.46
	46	0.76	0.43
	72	0.71	0.45
COS	25	0.75	0.46
	26	0.55	0.84
	27	0.67	0.52
	60	0.55	0.87
	61	0.51	0.95
LRS	75	0.60	0.80
	78	0.62	0.61
	38	0.65	0.63
	40	0.65	0.51
	71	0.59	0.63
SSS	77	0.60	0.83
	79	0.57	0.71
	81	0.44	0.78
	39	0.73	0.52
	76	0.57	0.94
LCS	29	0.76	0.44
	31	0.71	0.50
	34	0.61	0.58
	63	0.61	0.64
	32	0.65	0.73

Table 5.5: Correlations between new scales

	SSS	LRS	COS	LCS	GQS
SSS					
LRS	1.00				
COS	0.71	0.80			
LCS	0.68	0.65	0.63		
GQS	0.57	0.67	0.65	0.68	
IMS	0.58	0.69	0.70	0.66	0.98

5.4 Structural models covarying new with current CEQ scales

In addition to testing relationships within and between the new scales, it was important to also examine how six new scales related to the five current CEQ scales is important. A factor model was examined which treated each new scale as a single factor and allowed covariation between them. The inter-scale correlations are given below. In this model, each scale was covaried with all others.

Table 5.6: Interscale correlation patterns

	SSS	LRS	COS	LCS	GQS	IMS	GTS	GSS	CGS	AWS
SSS										
LRS	1.00									
COS	0.71	0.80								
LCS	0.68	0.65	0.63							
GQS	0.57	0.67	0.65	0.68						
IMS	0.58	0.69	0.70	0.66	0.98					
GTS	0.54	0.60	0.71	0.70	0.60	0.65				
GSS	0.47	0.54	0.52	0.61	0.83	0.75	0.58			
CGS	0.58	0.64	0.72	0.56	0.55	0.59	0.69	0.51		
AWS	0.23	0.23	0.39	0.15	0.19	0.20	0.31	0.07	0.39	
AAS	0.20	0.31	0.36	0.24	0.42	0.45	0.43	0.34	0.37	0.37

The shaded part of this table, presents the correlations between the old and new scales and is illustrated in the figure below.

Figure 5.1: Graph of information in shaded portion of table above

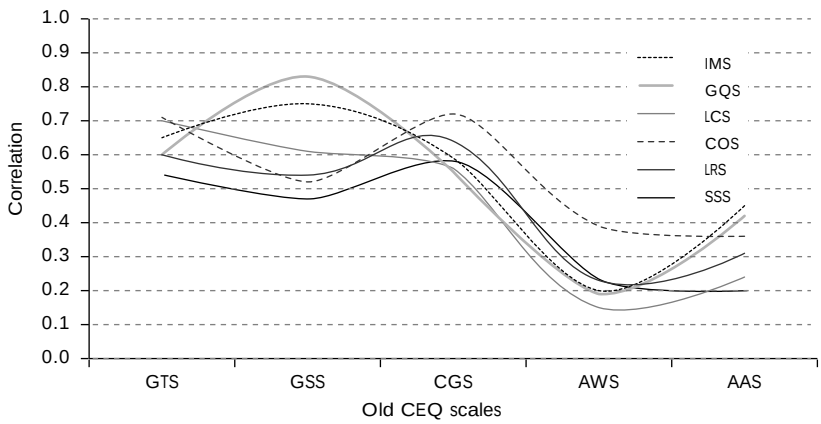


Figure 5.1 illustrates several trends among the inter-scale correlation. It also displays the differential behaviours of the new scales but in particular can be used to ascertain whether the current CEQ and new scales may contaminate

each other. It is important to note that despite the high correlation between some new and current CEQ scales shown above, the salient consideration is that the presence of a strong statistical relationship between two variables does not imply that they are measuring the same thing. Rather, it implies that the things being measured fluctuate sympathetically. Thus, the relationship of new to current CEQ scales does not imply redundancy but, conversely, may even imply congruence. The generally low overall correlation between the scale sets suggests independence.

5.5 Scale reliabilities at aggregate levels

Several multilevel regression analyses were conducted. The purpose of these analyses was to examine variance components at levels within the sample. The rationale for this was the argument raised by Marsh and Rowe (1999) that the CEQ and like instruments were used to make decisions at a level of institution or at the level of a field of study or faculty within an institution, but were not made at the student level, despite the fact that the data was collected at that level.

Hence it would be optimal to examine a three-level model with students nested within universities which are in turn nested within fields of study. This hierarchical structure was chosen because comparisons are made for like faculties between universities, but not for between faculties within universities. In fact it was not possible to conduct three level models because of the structure of the data file. Not all fields of study are present for all universities and the purposeful sampling process described in Chapter 1 ensure that fields of study were represented in the data regardless of university representation. This was reinforced by the fact that not all universities were participating in the study. Consequently, a series of two level analyses were conducted within sub samples for fields of study.

The dependent variable was each student's measure of satisfaction on each of the scales, including the CEQ scales. χ^2 tests of differences between likelihood ratio statistics for a series of two-level models were conducted. The levels in this instance were universities, fields of study and students.

The multilevel model used in the analysis was $y_{ij} = \beta_{0ijx0}$, where $\beta_{0ij} = \beta_0 + u_{0j} + e_{0ij}$. Reliability estimates were obtained using Bosker and Snijders (1999) approach where reliability of an aggregated variable is estimated using the following for reliability of aggregated data λ_j:

$$\lambda_j = \frac{n\rho}{1 + (n-1)\,\rho}$$

where: ρ = intraclass correlation and n = group size (variable in our case). In this case the mean group size was used as the estimate of cluster size. Two Fields of Study (BFOS) provided sufficient cases to demonstrate the process and to estimate the reliability.

Table 5.7: Variance and reliability estimates at aggregated levels

BFOS	Scale	FIXED		RANDOM							
		Random intercept		Variance between UNIVERSITIES (σ_{u^2})			Variance between STUDENTS (σ_{e^2})			Total variance (σ_{T^2})	
		β_0	(SE)	σ_{u^2}	(SE)	%	σ_{e^2}	(SE)	%	σ_{T^2} (SE)	λ_j
BFOS 7	SSS	0.311	0.093	0.055	0.041	2.893	1.846	0.112	97.107	1.901	0.61
Humanities	LRS	0.670	0.103	0.059	0.050	2.178	2.650	0.160	97.822	2.709	0.54
and Social	COS	0.173	0.071	0.017	0.023	0.895	1.882	0.114	99.105	1.899	0.32
Sciences	LCS	–0.007	0.159	0.217	0.122	6.626	3.058	0.185	93.374	3.275	0.79
(n*≅ 57)	GQS	1.419	0.148	0.179	1.105	5.492	3.080	0.186	94.508	3.259	0.75
	IMS	1.822	0.195	0.309	0.184	5.399	5.414	0.328	94.601	5.723	0.75
BFOS 9	LRS	0.451	0.063	0.005	0.015	0.248	2.010	0.114	99.752	2.015	0.10
Sciences	SSS	0.231	0.061	0.008	0.015	0.491	1.622	0.092	99.509	1.630	0.19
(n*≅ 46)	COS	0.064	0.084	0.038	0.031	2.262	1.642	0.093	97.738	1.680	0.52
	LCS	–0.019	0.123	0.099	0.069	3.534	2.702	0.153	96.466	2.801	0.63
	GQS	1.306	0.140	0.142	0.091	4.749	2.848	0.162	95.251	2.990	0.70
	IMS	1.607	0.227	0.434	0.242	8.052	4.956	0.281	91.948	5.390	0.80

(n*) is the average cluster size across universities.

Sampling for this study did not directly build in the capacity to examine between institutional differences. However because of the PREQ experience, some examination of this was thought to be necessary. The aggregated reliabilities of the scales varied considerably at the level of university within the broad field of study. Certainly the reliability of the aggregated data is greater than any of the estimates provided in Table 5.8. What is clear is that most of the variance is at the student level for the limited field of study in which this examination was possible. The figures above vary between 92 per cent and 99 per cent. Very little is explained by differences between universities within those fields of study. (Though it may be the case that the much larger number of responses in the full CEQ, leading to an increase in statistical power, might impact on reliabilities). However little is known about the relativity of reliabilities at these levels and it is not possible to comment on these values. Sampling emphasised student level data and hence the variance is identifiable at student level. Very little can be explained by higher levels and this in turn means that the reliability must be low at these levels, at least using the Bosker and Snijders (1999) approach. If the reliability becomes the proportion of systematic variance taken as a fraction of the available variance it would be quite high, but it is still not clear what it might mean. In the context of the full CEQ, rather than the sample framework used in this study, we need to consider the appropriateness of the multilevel modelling

approach. For example, institutional comparisons using the full CEQ are conditioned on institutions which represent the complete population rather than the sample of institutions taken in this study. Further work in this area at a methodological level is needed to ascertain its importance and impact on decision making related to the CEQ data.

References

Adams, R. J. & Khoo, S. T. (1994), *QUEST: The Interactive Test Analysis System*, ACER, Melbourne.

Ainley, J. & Long, M. (1994), *The 1995 CEQ: An Interim Report*, ACER, Melbourne.

Ainley, J. & Long, M. (1994), *The Course Experience Survey 1992 Graduates*, Australian Government Publishing Service, Canberra.

Bollen, K. A. (1989), *Testing Structural Equation Models*, Wiley, NY.

Bollen, K. A. & Long, J. S. (1993), *Testing Structural Equation Models*, Sage Publications, NY.

Byrne, B. M. (1994), 'Testing for the Factorial Validity, Replication and Invariance of a Measurement Instrument: A Paradigmatic Application Based on the Maslach Burnout Inventory', *Multivariate Behavioral Research*, Vol. 29, No. 3. pp. 289–311.

Byrne, Shavelson & Muthén, (1989), 'Testing for the Equivalence of Factor Covariance and Mean Structures: The Issue of Partial Measurement Invariance', *Psychological Bulletin*, Vol. 103, No. 3, pp. 456–66.

Cashin, W. E. & Downey, R. G. (1992), 'Using Global Student Rating Items for Summative Evaluation', *Journal of Educational Psychology*, Vol. 84, No. 4, pp. 563–72.

Chang, C.H. (1996). 'Finding Two Dimensions in MMPI-2 Depression', *Structural-Equation-Modeling*, Vol. 3, No. 1, pp 41–49.

Curtis, D. (1999), *The 1996 Course Experience Questionnaire: A Re-Analysis*, Research Paper, School of Education, Flinders University, South Australia.

DETYA (1998), *Selected Higher Education Student Statistics*, Australian Government Publishing Service, Canberra.

Eley, M. G. (undated), *The Course Experience Questionnaire: Some Effects of Altered Question Formats on the CEQ's Effectiveness as an Instrument*, Centre for Higher Education Development, Monash University, Melbourne.

Elphinstone, L. (1989), *The Development of the Course Experience Questionnaire*, M.Ed Thesis, Centre for the Study of Higher Education, The University of Melbourne.

Entwistle, N. J. & Ramsden, P. (1983), *Understanding Student Learning*, Croom Helm, London.

GCCA (1999), *1998 Graduate Destination Survey*, Graduate Careers Council of Australia Limited, Victoria.

Goekoop, J.G. and Zwinderman, A.H. (1994). 'Multidimensional hierarchic ordering of psychopathology: Rasch analysis in factor-analytic dimensions', *Acta-Psychiatrica-Scandinavica*, Vol 90. No, 6, pp 399–404.

Green. K.E. (1996). 'Dimensional Analyses of Complex Data', *Structural-Equation-Modeling*, Vol. 3, No. 1, pp 50–61.

Griffin, P. (2000), *Measuring Achievement Using Sub Tests From a Common Item Pool: A Cross National Application of the Rasch Model*, (In Press), Paris, IIEP.

Hand, T., Trembath K. & The University of New South Wales (1999), *The Course Experience Questionnaire Symposium 1998*, DETYA, Higher Education Division, 99/2, Canberra.

Johnson, T. (1999), *1998 Course Experience Questionnaire: A Report Prepared for the GCCA*, Graduate Careers Council of Australia Limited, Parkville.

Johnson, T. (1998), *The Course Experience Questionnaire: a Report Prepared for the GCCA*, Graduate Careers Council of Australia Limited, Parkville.

Johnson, T. (1997), *The 1996 CEQ*, Graduate Careers Council of Australia Limited, Parkville.

Karmel, T., Aungles, P. & Andrews, L. (1998), 'Presentation of the CEQ', paper delivered to the CEQ Symposium on 30 September 1998 at University of New South Wales.

Karmel, T., Aungles, P. & Andrews, L. (1998), 'Presentation of the Course Experience Questionnaire (CEQ),' paper delivered to the CEQ Symposium, University of New South Wales.

Linacre, J. M. (1998), 'Rasch First or Factor First?', *Rasch Measurement Transactions*_Vol. 11, No. 4, p. 603.

Long, M. & Johnson, T. G. (1997), *Influences on the CEQ Scales*, Australian Government Publishing Service, Canberra.

Long, M. & Johnson, T. (1997), *Influences on the Course Experience Questionnaire Scales*, Evaluations and Investigations Program, DETYA, Canberra.

Magin, D. (1991). 'Issues relating to the use of the CEQ as a performance indicator', 'Papers presented at the sixteenth annual conference of the Higher Education Research Society of Australasia, held at Griffith University, Brisbane, 6th–9th July, 1990' Ross, B. (ed.)., pp 363–369. NSW: Higher Education Research and Development Society of Australasia (HERDSA).

Marsh, H. W. (1987), 'Students' Evaluations of University Teaching', *International Journal of Educational Research*, Vol. 1, No. 25, pp. 13–88.

Marsh, H. W. & Overall, J. U. (1981), 'The Relative Influence of Course Level, Course Type and Instructor on Students' Evaluations of College Training', *American Educational Research Journal* , Vol. 18, No. 1, pp. 103–112.

McInnis, C. (1997), Defining and Assessing the Student Experience in the Quality Management Process, *Tertiary Education and Management*, No. 3, pp. 63–71.

McInnis, C. & James, R. (1995), *First Year on Campus*, Australian Government Publishing Service, Canberra.

McInnis, C., Baldwin, G. & James, R. (1998), *Measuring Dimensions of the Student Experience Beyond Instruction*, CEQ Symposium, 29–30 September 1998, University of New South Wales.

Pace, C. (1979), *Measuring Outcomes of College: Fifty Years of Findings and Recommendations for the Future*, Jossey-Bass, San Francisco.

Pascarella, E. (1985), College Environmental Influences on Learning and Cognitive Development: A Critical Review and Synthesis. J. Smart (ed), *Higher Education: Handbook of Theory and Research* , Vol. 1, Agathon, New York.

Pascarella, E. & Terenzini, P. (1991), *How College Affects Students*, Jossey-Bass, San Francisco.

Pascarella, E. & Terenzini, P. (1998), Studying College Students in the 21st Century: Meeting New Challenges, *Review of Higher Education*, Vol. 21, pp. 151–165.

Pentz, M. A. & Chou, C. P. (1994), 'Measurement Invariance in Longitudinal Clinical Research Assuming Change from Development and Intervention', *Journal of Consulting and Clinical Psychology*, Vol. 62, No. 3, pp. 450–62.

Ramsden, P. (1991), 'A Performance Indicator of Teaching Quality in Higher Education: the Course Experience Questionnaire', *Studies in Higher Education*, Vol. 16, pp. 129–149.

Ramsden, P. (1991a), 'Report on the CEQ Trial', in R. Linke *Performance Indicators in Higher Education*, Vol 2, Australian Government Publishing Service, Canberra.

Ramsden, P. (1991b), 'A Performance Indicator of Teaching Quality in Higher Education: the Course Experience Questionnaire', *Studies in Higher Education*, Vol. 16, No. 2, pp. 129–150.

Ramsden, P. (1992), *Learning to Teach in Higher Education*, Routledge, London.

Ramsden, P. (1996), 'The Validity and Future of the CEQ', paper presented on 3–4 October at the *AVCC CEQ Symposium* at Griffith University, Queensland, Australia.

Rasch, G. (1960). *Probabilistic Models for Some Intelligence and Attainment Tests.* Chicago: MESA Press.

Richardson, J. T. E. (1994), A British Evaluation of the Course Experience Questionnaire, *Studies in Higher Education*, Vol. 19, pp. 59–68.

Schumacker, R. E. & Linacre, J. M. (1996), 'Factor Analysis and Rasch', *Rasch Measurement Transactions*, p. 470.

Smith, R.M. (1996). 'A Comparison of Methods for Determining Dimensionality in Rasch Measurement', *Structural-Equation-Modeling*, Vol. 3, No. 1, pp 25–40.

Terenzini, P. & Wright, T. (1987), Students' Personal Growth During the First Two Years of College, *Review of Higher Education*, Vol. 10, No. 2, pp. 259–271.

Tinto, V. (1987), *Leaving College: Rethinking the Causes and Cures of Student Attrition*.

Treloar, D. W. G. (1994), *The Course Experience Questionnaire: Appreciation and Critique*, The University of Western Australia.

Waugh, R. F. (1998), The Course Experience Questionnaire: a Rasch Measurement Model Analysis, *Higher Education Research and Development*, Vol. 17, pp. 45–63.

Wilson, K. L., Lizzio, A. & Ramsden, P. (1997), The Development, Validation and Application of the Course Experience Questionnaire, *Studies in Higher Education*, Vol. 22, pp. 33–52.

Wright, B. D. (1996), 'Comparing Rasch Measurement and Factor Analysis', *Structural Equation Modelling*, Vol. 3, No. 1. pp. 3–24.

Wright, B. D. & Masters, G. N. (1983), *Rating Scale Analysis*, MESA Press, Chicago.

Yorke, M. (1996), *Indicators of Programme Quality*, Higher Education Quality Council, London.