THE UNIVERSITY OF ADELAIDE AUSTRALIA

Course and Teacher SELT Questionnaires

Validity, Reliability and Measurement Properties

David D Curtis

Francisco Ben

School of Education, The University of Adelaide

March 2009

Contents

Executive Summary ..... **vi**
Data and methods vi
Key findings vi
Recommendations vii
1 Introduction **1**
2 Data and Methods **2**
The SELT questionnaires 2
Summary of SELT responses 3
Analytic approaches **4**
Descriptive and exploratory analysis 4
Confirmatory factor analysis 4
Rasch analysis 6
3 Results **8**
Course SELT **8**
Descriptive summary 8
Exploratory analysis 9
Confirmatory factor analysis 10
Rasch analysis 11
Teacher SELT **16**
Descriptive summary 17
Exploratory analysis 17
Confirmatory factor analysis 18
Rasch analysis 19
4 Supplementary Analyses **23**
What is the relationship between workload and perceived course quality? 23
By how much do perceptions of course quality vary between courses? 24
By how much do perceptions of teaching quality vary between teachers? 25

5 Conclusions **27**
Key findings **27**
The Course SELT 27
The Teacher SELT 29
Implications and recommendations **29**
6 References **31**

List of Tables

Table 1: Text for the 15 standard Likert-response items for the Course SELT 2
Table 2: Text for the seven standard Teacher SELT items 3
Table 3: Course SELT responses by academic year and faculty, school and discipline 3
Table 4: Teacher SELT responses by academic year and faculty, school and discipline 4
Table 5: Response frequencies for the Course SELT items 8
Table 6: Results of scale reliability analysis for the Course SELT 9
Table 7: Factor loadings for the 15 Course SELT items 9
Table 8: Factor loadings for the Course SELT items after three CFA runs 10
Table 9: Summary of CFA results on Course SELT single factor model. 11
Table 10: Item fit indices for the Course SELT instrument 14
Table 11: Response frequencies for the Teacher SELT items 17
Table 12: Results of scale reliability analysis for the Teacher SELT 17
Table 13: Factor loadings for the Teacher SELT 18
Table 14: Factor loadings for the Teacher SELT items 18
Table 15: Summary of CFA results on Teacher SELT single factor model 18
Table 16: Item fit indices for the Teacher SELT 20

List of Figures

Figure 1: Structures of the single factor models for the Course SELT and Teacher SELT questionnaires 5
Figure 2: 'Cave' plot of item and person distributions for the Course SELT 13
Figure 3: Response distribution map for the Course SELT 16
Figure 4: 'Cave' plot of item and person distributions for the Teacher SELT 20
Figure 5: Response distribution map for the Teacher SELT 21
Figure 6: Relationship between perceived course quality and perceptions of course workload 23
Figure 7: Mean values and confidence intervals of perceived course quality by course 25
Figure 8: Mean values and confidence intervals of perceived teaching quality by teacher 26

Executive Summary

Student perceptions of the quality of their experience of the learning and teaching environment at the University of Adelaide are assessed through the administration of survey questionnaires – the Student Experience of Learning and Teaching (SELT). Different versions of the SELT instrument are used for courses, teachers and programs. In this study, data collected from the administration of Course and Teacher SELTs in large first year classes have been analysed in order to evaluate the metric properties of these instruments.

One of the principles affirmed in the SELT Policy is to "improve student learning outcomes" and to use SELT processes to contribute to this goal. By emphasising students' experiences of learning, SELT processes seek to enhance the quality of learning and, therefore, the outcomes of that learning. The SELT surveys, therefore, have a formative role in teaching practices within the university. For SELT surveys to provide useful information about their contribution to learning outcomes, the SELT forms must lead to coherent data that have a demonstrable relationship with valued learning outcomes. We analyse data sets collected using current course and teacher SELT questionnaires to investigate the coherence of the data and through those data, the reliability and precision of the questionnaires.

Teacher SELT data are used in support of applications for tenure and promotion and for the award of prizes and grants. They have, therefore, a summative role in contributing to judgments of quality, although they are certainly not the only source of information used. It is important to demonstrate that the teacher SELT questionnaire is capable of providing the level of detail required for these judgments.

The research questions advanced prove a framework for investigating the Course and Teacher SELT instruments.

- Do the items in the SELT course and teacher questionnaires function as expected?
- Do the data gathered from administrations of the course and teacher SELT questionnaires adequately cover the range of respondent views and do they provide sufficiently reliable and precise estimates of perceived teaching quality and learner engagement for the purposes to which these data are put?
- Do the SELT questionnaires provide meaningful and objective evaluation of quality of students' experiences of teaching and learning?
- Can aberrant responses to the SELT forms be detected, and if so, how can these responses be managed?

Data and methods

Data from over 8,000 responses to the Course SELT questionnaire and over 17,000 to the Teacher SELT questionnaire in large first-year classes at the University of Adelaide were analysed. A range of analytic techniques, including basic descriptive statistics, exploratory factor analysis, confirmatory factor analysis and Rasch analysis were undertaken on these responses.

Key findings

Both the Course and Teacher SELT instruments have quite reasonable measurement properties. Decisions based on them in their present form are soundly based.

In both the Course and Teacher SELTS, there is scope for improvement. The Course SELT includes an item on workload. This is not related to other items in the instrument. Its inclusion does not compromise the Course SELT, but it does not contribute to the measurement of perceived course quality.

Two items, one about the absence of discrimination and the other about considering students backgrounds, are related to each other. They do not show a strong relationship with other items in the instrument, but they may have particular salience for sub-groups of the student body. They are worthy items and would be more meaningful if additional information were available about the students who respond, perhaps information on gender or country of birth.

The Teacher SELT is a brief but informative instrument. All seven items cohere well and contribute to the measurement of students' perceptions of the quality of teaching. The construct is not measured comprehensively. We believe that the inclusion of some additional items could improve this instrument.

Both instruments, especially the Teacher SELT, could be better targeted to the high esteem that students have for the quality of courses and teaching they experience at the University of Adelaide. Many students choose very favourable responses categories and few choose the unfavourable ones. For the instrument to the maximum information, a more symmetrical distribution of responses is desirable. It is very likely that having fewer response categories would yield similar information to that obtained from the current instrument, but a better solution is to include items that students are less likely to endorse so strongly.

We find that relatively few students (<5%) provide inconsistent responses to these instruments. This suggests that students take the evaluation of courses and teaching seriously and their responses make a substantial contribution to the University's quality improvement processes.

Recommendations

1 Data collected from the administration of the Course and Teacher SELT instruments be analysed using a strong measurement approach. The partial credit Rasch measurement model is suggested.

The use of a sound measurement-based approach to the analysis of Course and Teacher SELT data will enable improvements to be made to the current instruments over time while preserving the meaning of scaled scores. If questions are added that are more difficult to endorse favourably, average raw scores will decline. The measurement approach will take into account the difficulty of the questions in scaling student responses.

This approach will also support the fair, valid and reliable comparison of courses and teaching over time.

2 Perceived course and teaching quality be reported on a defined scale with a mean of 500 and a standard deviation of 100 units.

The choice of origin (zero point) and the size of a measurement unit are matters of preference, similar to choosing between the Fahrenheit or Celsius scales for reporting temperature. In reporting academic achievement, it is common to use scales with a mean of 500 and a standard deviation of 100 units. This is done, for example, with the Scholastic Aptitude Test (SAT) and in reporting literacy achievement in the Programme for International Student Achievement (PISA). While this scale is a departure from current practice and users of SELT data would need to acclimatise to the scale, it would soon become common currency in teaching discourse.

3 Additional demographic information, specifically respondent gender and main language background (English/other) be sought on the Course and Teacher SELT instruments.

The collection of additional information about student respondents, in particular gender and language background, would assist in the interpretation of the results of analyses and contribute to quality improvement within the University. Some items in the Course SELT appear to show particular response patterns that might be consistent with minority groups of students having different views from the majority. The additional data would enable the identification of groups who may have different needs or expectations from the majority. Teaching approaches could be adapted to meet the needs of diverse groups if the hypothesised differences are confirmed.

In small classes, students may be concerned about the possibility of being identified. Strong assurances that this is not the intention accompanied by an explanation for the data may alleviate such concerns.

1 Introduction

Student perceptions of the quality of their experience of the learning and teaching environment at the University of Adelaide are evaluated through the administration of survey questionnaires – the Student Experience of Learning and Teaching (SELT). Different versions of the SELT instrument are used for courses, teachers and programs. In this study, data collected from the administration of Course and Teacher SELTs in large first year classes have been analysed in order to evaluate the metric properties of these instruments.

The Centre for Learning and Professional Development (CLPD) has responsibility for the "design, delivery, evaluation and improvement of SELT forms" (University of Adelaide, SELT Policy, p. 3). One of the goals of the CLPD is to "implement appropriate methods for the evaluation of student learning and staff teaching using methods that are informed by educational research" (http://www.adelaide.edu.au/clpd/evaluation/seltsystem.html). This report appraises the metric properties of the Course and Teacher SELT questionnaires.

Teacher SELT data are used in support of applications for tenure and promotion and for the award of prizes and grants. They are not the only information used for these purposes. They have a summative role in supporting judgments of teaching quality. We investigated the extent to which Teacher SELT data can discriminate levels of teaching quality as perceived by students.

The research questions advanced prove a framework for investigating the Course and Teacher SELT instruments in large first-year classes.

- Do the items in the Course and Teacher SELT questionnaires function as expected?
- Do the data gathered from administrations of the Course and Teacher SELT questionnaires adequately cover the range of respondent views? Do they provide sufficiently reliable and precise estimates of perceived teaching quality and learner engagement for the purposes to which these data are put?
- Do the SELT questionnaires provide meaningful and objective evaluation of quality of students' experiences of teaching and learning?
- Can aberrant responses to the SELT forms be detected, and if so, how can these responses be managed?

The remainder of the report is structured to provide:

- an account of the data sources and analytic methods used (Section 2);
- results of analyses (Section 3), including for both the Course and Teacher SELTs:
  - presentation of basic descriptive statistics;
  - the results of the confirmatory factor modelling;
  - the results of the Rasch analyses;
  - a summary of the main results;
- supplementary analyses are reported (Section 4) to investigate responses to one misfitting item in the Course SELT and to explore the possibility of future analyses that could follow from a strong measurement approach to the Course and Teacher SELT instruments; and
- a set of implications that flow from these analyses with suggestions for action (Section 4).

Summary, rather than comprehensive, results are presented in the body of the report. Tables with detailed results of analyses are included as appendices.

2 Data and Methods

In this section, the data sources and methods used for their analysis are outlined.

Approval for the project to proceed was granted by the Research Ethics Committee. Data were extracted from Course and Teacher SELT databases managed by the CLPD. All information identifying the faculty, school, discipline and teachers were removed from the extracted data. All data are from large first year courses from the years 2006 to 2008. The 'confidentialised' data were provided to the investigators by the CLPD.

The SELT questionnaires

The Course SELT comprises 15 standard items. These items are shown in Table 1.

Table 1: Text for the 15 standard Likert-response items for the Course SELT

Item no.	Item text
1	Overall, how would you rate the workload in this course?
2	Overall, I am satisfied with the quality of this course.
3	The course stimulates my enthusiasm for further learning.
4	I feel part of a group committed to learning.
5	It is made clear what is expected of me.
6	I receive adequate feedback on my work.
7	I am motivated to learn in this course.
8	The assessment allows me to demonstrate what I understand.
9	This course helps me develop my thinking skills (eg. problem solving, analysis).
10	The learning resources (eg. handouts, web resources) are valuable for my understanding of the course.
11	I am satisfied with the course information provided (eg. course outlines, assessment details, timetables).
12	The learning environment is free from discrimination.
13	The learning environment takes into account the students' backgrounds.
14	My ability to work independently is being increased.
15	I understand the concepts presented in this course.

Each item has seven response options, with a further 'Not applicable' (N/A) option. With the exception of the first item, the extreme options are labelled 'Strongly agree' and 'Strongly disagree' and the middle option is labelled 'Neutral'. The corresponding labels for the first item are 'Very heavy', 'Very light' and 'Reasonable'.

The teacher SELT comprises nine standard items. Seven of these are select items for which students choose one of the provided options and the final two are supply items for which student enter a free text response. Only responses to the seven select items were analysed.

The seven select items in the Teacher SELT are shown in Table 2.

Table 2: Text for the seven standard Teacher SELT items

Item no.	Item text
1	All things considered, how would you rate the effectiveness of [teacher] as a university teacher.
2	[Teacher] is well organised.
3	[Teacher] shows concern for students.
4	[Teacher] shows enthusiasm for encouraging student learning.
5	[Teacher] encourages student participation.
6	[Teacher] stimulates my interest in learning in this course.
7	[Teacher] gives clear explanations.

In addition to a not applicable (N/A) option, these items have seven response options. Response options for the first item are labelled 'Outstanding' and 'Very poor' (extreme options) and 'Reasonable' (middle option). The remaining items are labelled 'Strongly agree' and 'Strongly disagree' (extreme options) and 'Undecided' (middle option).

Summary of SELT responses

Summaries of the SELT responses are shown below; Course SELTs in Table 3 and Teacher SELTs in Table 4. Data from 8,269 Course SELT questionnaires from five faculties, 14 schools and 17 disciplines from the 2006, 2007 and 2008 academic years were available for analysis. Similarly, 17,905 Teacher SELT responses from five faculties, 18 schools and 23 disciplines were available.

Table 3: Course SELT responses by academic year and faculty, school and discipline

Faculty	School	Discipline	Academic year			Total
			2006	2007	2008	
F1	S1	D1	289	345		634
	S2	D2	121	191	151	463
F2	S5	D7	91		111	202
	S6	D8	208	153	189	550
F3	S7	D9		134	80	214
	S8	D10			66	66
		D11	119			119
	S9	D12		98	82	180
		D13			118	118
F4	S11	D15		211		211
	S12	D16		169	66	235
	S13	D17	353	291		644
	S14	D18	846	885	787	2518
F5	S15	D19		114		114
	S16	D20	505	356	242	1103
		D21		116	130	246
	S18	D27	228	424		652
Totals			2760	3487	2022	8269

Table 4: Teacher SELT responses by academic year and faculty, school and discipline

Faculty	School	Discipline	Academic year			Total
			2006	2007	2008	
F1	S1	D1	247			247
	S2	D2		467	251	718
	S3	D3	87	139		226
		D4	490	194	257	941
F2	S4	D6		121		121
	S5	D7	174	341		515
	S6	D8	769	443	400	1612
F3	S7	D9		81		81
	S8	D11	138	122	124	384
	S9	D12		177	80	257
		D13	72		116	188
F4	S10	D14			412	412
	S11	D15		104		104
	S12	D16	229	139	155	523
	S13	D17	1005	1117		2122
	S14	D18	1191	1715	1883	4789
F5	S15	D19		109		109
	S16	D20	716	1468	1175	3359
		D21	171	158	136	465
	S17	D22			240	240
		D23	73			73
	S18	D26	197			197
		D28		126	96	222
Total			5559	7021	5325	17905

Analytic approaches

After initial data checks, a series of descriptive and exploratory analyses were undertaken. The two main analytic methods were confirmatory factor analysis and Rasch analysis. Confirmatory factor analysis is used to establish the fit of items to a common hypothesised factor. Rasch analysis is used to verify that items cohere to measure an underlying quality construct and to generate measures for individuals. The method is also used to assess the consistency of individual's responses. In studies of students' responses to the Course Experience Questionnaire, approximately 20 per cent of responses were found to be aberrant (Curtis & Keeves, 2000).

Descriptive and exploratory analysis

Basic descriptive summaries of SELT data are presented so that comparisons can be made with existing reports arising from the administration of SELT questionnaires.

Confirmatory factor analysis

Confirmatory Factor Analysis (CFA) was undertaken to verify the factor structure of the set of observed variables in both the Course and Teacher SELT questionnaires. It was also used to confirm the hypothesised relationships between the observed variables and their latent constructs. Linear Structural Relations (LISREL) (Joreskog & Sorbom, 2007) was used to carry out CFA. Softwares such as the Analysis of Moment Structures (AMOS) (Arbuckle, 2007) and MPlus (Muthen & Muthen, 2006) were also used, but only to check that the parameters and fit indices were estimated similarly by the various programs. There were some differences, as some programs, e.g. AMOS, make assumptions about the scale on which

variables are measured. We report the results obtained using LISREL (Joreskog & Sorbom, 2007). CFA is a multivariate procedure that uses a structural equation modelling (SEM) technique that explicitly provides estimates of error variance parameters (Byrne, 2001).

CFA was carried out to examine the structures of the SELT instruments because EFA is rather limited in its capability to explore the more involved structures that may be required to represent complex constructs. Curtis (2004, p.187) pointed out that:

> Tools such as exploratory factor analysis are limited in the extent to which they are able to probe these structures. **Further, for a construct to be compatible with simple measurement – that is, to be able to report a single quantitative score that truly reflects a level of particular construct – the structure of the construct must reflect ultimately a single underlying factor.** (Emphasis added)

Several models that are compatible with a single measurement can be tested with CFA to see whether they are consistent with the data. However, testing these models was not the aim of the study. In addition, the Course and Teacher SELT instruments only have 15 and 7 items, respectively – something that may be considered as 'few' compared to similar instruments. Furthermore, it is expected that these items are all load directly onto a single latent factor. Thus, the only model tested in this study was the single factor model (Figure 1). Results of the CFA are presented in Section 3.

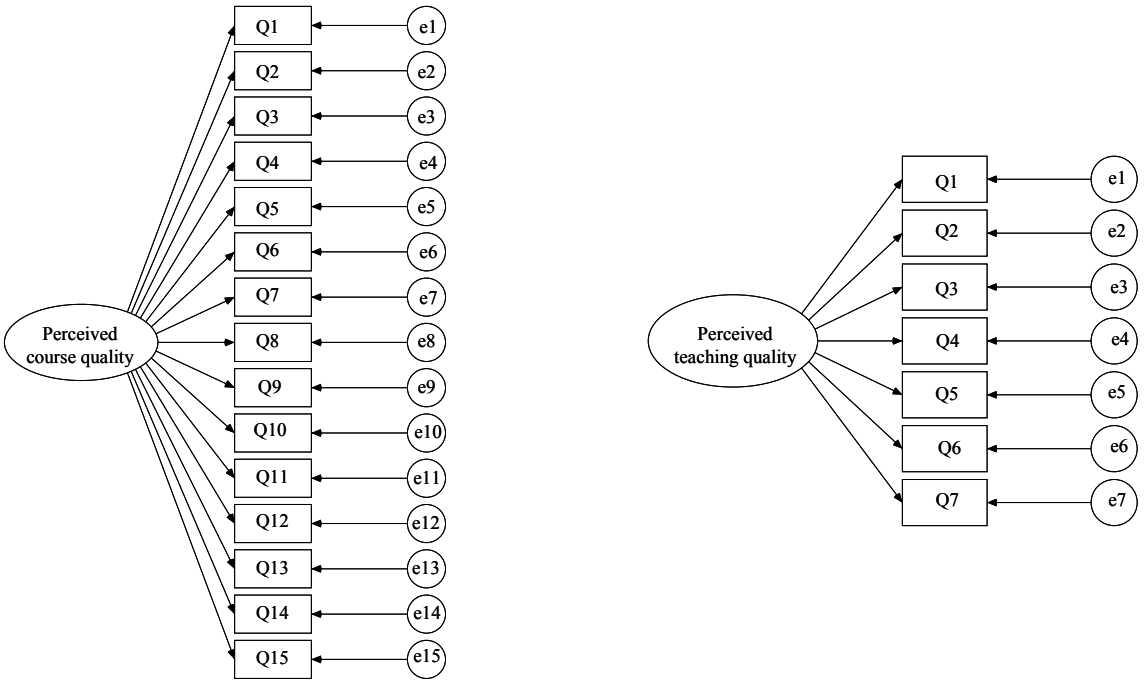

Figure 1: Structures of the single factor models for the Course SELT and Teacher SELT questionnaires

Much social science research, according to Keeves and Masters (1999), involves constructs that are often complex, multi-faceted, multivariate and multi-level. A conventional psychometric approach to scale construction involves operationalising constructs using very narrowly delimited items. An alternative is to recognise that complex constructs are necessarily multi-faceted. Their comprehensive measurement would involve developing instruments that included sets of items for each facet of the construct. This would lead to long instruments, and in the context of evaluating courses and teaching, would be onerous for respondents and would have poor response rates.

Rasch analysis

Responses to Likert items are often scored by assigning '1' to the least valued response, e.g. 'strongly disagree', and the highest score, perhaps '7', to the most favourable response, e.g. 'strongly agree'. Other schemes for scoring responses are also used. In analyses of the Course Experience Questionnaire (CEQ), it has been common practice to score the five response options -100, -50, 0, +50 and +100 (Johnson, 1997). In summarising responses, it is common to report a mean response score. For example, the 'mean' response score to the Course SELT item 'Overall, I am satisfied with the quality of this course' is 5.11. Taking the mean of the scored responses implies that the scores assigned to the responses form an interval scale. That is, the difference between the judgments 'strongly disagree' and 'disagree' and between 'neutral' and 'agree' are equal and represent a constant increase in the judgement of quality. However, the options offered to survey respondents do not lie on an interval scale: they are ordinal. We can say that agreeing with a proposition is better than being neutral, but we cannot say that this difference is the same as the difference between the 'strongly disagree' and 'disagree' responses. It is necessary to use methods that recognise the ordinal character of the data, and the Rasch measurement model does this (Harwell & Gatti, 2001).

Scoring responses to test items and to survey responses has challenged educational and social science researchers for much of the past 100 years. Rasch (1960, 1980) proposed a workable solution to this problem and several variants to this method have been developed. The basic Rasch formulation fits item difficulty and person ability parameters to a measurement model for responses to dichotomous items such as right or wrong answers to test questions. The Rasch method models the difference between the ability of the test-taker and the difficulty of test items. The probability of a correct response is a logistic function of the difference between the person ability and the item difficulty. For Likert-response items, two extensions of the basic Rasch formulations have been developed – the rating scale model (Andrich, 1978) and partial credit model (Masters, 1982). In this investigation, we use the partial credit model. In this model, in addition to estimating parameters for individual ability, parameters are estimated for the level of demand in crossing a threshold to agree with the next response category for each item. The threshold is the point at which a person is likely to switch from endorsing one option and choosing the next, e.g. choosing 'agree' instead of 'neutral'. As there are seven response options, there are six thresholds between them.

The Rasch method is used to estimate measures of individuals on an underlying scale and to estimate the locations of item thresholds on the same scale. Measurement demands that responses by individuals to items conform to requirements such as additivity (Michell, 1997). In assessing the conformity of responses to the requirements of measurement, the Rasch model generates fit indicators. Both items and individuals may fail to conform to measurement, and such misfit is revealed by these indicators. Where items fail to conform to measurement requirements, those items are deleted from the analysis. Similarly, where individuals misfit, those individuals may be removed from the analysis, especially when the instrument is being calibrated, as misfitting responses contribute noise rather than information to the calibration process. Further, the estimates assigned to misfitting individuals may not have the meaning that is imputed to them by the estimate (Curtis & Boman, 2004).

In the Rasch analyses that we present in this report, we evaluate the scales as effective measures at three levels. We examine the scale as a whole looking at item and person reliability indices and at the appropriateness of the targeting of the set of items; we report on the fit of items and persons to the measurement scale; and we investigate the separation of the item thresholds (Curtis & Boman, 2007). In part, the Rasch measurement model provides similar information to that given by CFA, but it goes beyond that by providing information on the conformity to measurement requirements of individuals' responses, by yielding

individual estimates that can be used in other analyses, and, by examining thresholds, it gives diagnostic information that can be used to refine the instrument.

By fitting item and person parameters, the Rasch measurement model generates an interval scale from ordinal data. Both items and persons are located on this common scale. The constructs of interest in measurement studies rarely have natural measurement scales. This is not unusual in the physical sciences where, for example, temperature is reported on either the Celsius or Fahrenheit scales. These scales have different zero points and different units, one degree Celsius being 1.8 times the magnitude of one degree Fahrenheit. Measurements on the two scales can be compared through a simple arithmetic transformation. In Rasch measurement, the zero point is normally set at the mean of the items and the units (logits) arise from the logistic function in which item difficulties and person abilities are modelled. It is common to convert measurement units to a person mean of 500 with a standard deviation of 100 units. This is done, e.g. with the Scholastic Aptitude Test (SAT). This transformation is applied to the scales generated for the Course and Teacher SELTs in this report.

3 Results

Results of the analyses are presented first for the Course SELT and then for the Teacher SELT. For each questionnaire, descriptive summaries are presented, followed by exploratory analyses, confirmatory factor analysis and several results of Rasch analysis.

Course SELT

The Course SELT comprises 15 common Likert-response items. Academics may include a further four Likert items and one free text item. Data for extended questions not included in analysis.

Descriptive summary

Frequencies of the various response options to the Course SELT items are shown in Table 5. Relatively few students choose the lowest three options and option 5, on the affirmative side of 'Undecided', is the modal response category for most items. In this table, mean response values are presented. We have argued that this is not an appropriate statistic for ordinal data, but this is a common summary statistic for such data.

Table 5: Response frequencies for the Course SELT items

	Response option labels [a]									
Item	Strongly disgree	2	3	Un-decided	5	6	Strongly Agree	Not applic-able	Missing	Mean
1	89	139	509	3101	2395	1525	475	7	29	4.71
2	73	179	522	1257	2880	2626	708	9	15	5.11
3	212	383	851	1483	2408	2030	887	8	7	4.83
4	200	456	1049	2303	2358	1395	439	56	13	4.48
5	121	320	784	1400	2549	2212	859	13	11	4.94
6	413	685	1258	1796	1985	1467	564	88	13	4.34
7	253	502	862	1642	2485	1779	723	12	11	4.68
8	170	358	735	1579	2484	2127	744	48	24	4.86
9	94	234	590	1411	2599	2393	912	19	17	5.07
10	117	225	501	1190	2141	2524	1502	45	24	5.27
11	93	172	439	1041	2283	2688	1518	9	26	5.35
12	90	109	221	878	1370	2504	2918	148	31	5.78
13	131	170	398	1641	1912	2209	1484	285	39	5.21
14	94	166	423	1521	2438	2513	1037	49	28	5.16
15	194	335	645	1340	2485	2278	956	13	23	4.97

Notes: [a] Response labels are those used for items 2 to 15. For Item 1, the three labelled response options are 'Very heavy', 'Reasonable' and 'Very light'.
The text for these items is shown in Table 1.
Data were available for 8,269 Course SELT questionnaires.

Cronbach's alpha for the set of items was found to be 0.904 (see Table 6). This is quite high and suggests that the set of items cohere to form a useful scale. The first item, about workload, does not sit well with the remaining items. Its item-total correlation is very low at 0.02, and if it were removed, the value of Cronbach's alpha for the remaining items would be 0.915. This item appears to be unrelated to the remaining items. Two other items, 12 and 13,

have low item-total correlations (<0.50) and their removal would make a very marginal difference to the value of alpha, suggesting that they make a very modest contribution to the scale. These items ask about discrimination and students' backgrounds. Rasch analyses, presented below, enable the structure of these items to be evaluated in more detail.

Table 6: Results of scale reliability analysis for the Course SELT

Item no.	Scale mean if Item deleted	Scale variance if Item deleted	Corrected Item-total correlation	Cronbach's Alpha if Item deleted
1	70.20	168.08	0.02	0.915
2	69.80	148.31	0.71	0.894
3	70.07	143.66	0.71	0.894
4	70.41	147.38	0.65	0.896
5	69.97	147.18	0.65	0.896
6	70.56	146.61	0.56	0.900
7	70.23	141.82	0.76	0.891
8	70.05	146.04	0.67	0.895
9	69.84	147.54	0.68	0.895
10	69.64	148.98	0.59	0.898
11	69.56	148.73	0.63	0.897
12	69.14	153.65	0.46	0.903
13	69.68	151.87	0.49	0.902
14	69.73	150.00	0.61	0.897
15	69.94	146.68	0.63	0.897
Scale alpha				0.904

Exploratory analysis

Exploratory factor analyses were undertaken on the Course SELT data using Mplus (Muthen & Muthen, 2006) and the results of these analyses are shown in Table 7.

Table 7: Factor loadings for the 15 Course SELT items

Item	Number of factors					
	One	Two		Three		
1	0.012	-0.007	0.047	-0.005	0.016	0.064
2	0.783	0.784	0.159	0.763	0.258	-0.043
3	0.783	0.846	0.037	0.848	0.226	-0.288
4	0.688	0.683	0.150	0.663	0.221	0.000
5	0.704	0.654	0.252	0.650	0.189	0.263
6	0.609	0.590	0.163	0.588	0.117	0.205
7	0.831	0.875	0.085	0.856	0.214	-0.102
8	0.730	0.703	0.206	0.706	0.136	0.267
9	0.726	0.678	0.252	0.670	0.209	0.222
10	0.640	0.522	0.407	0.513	0.294	0.381
11	0.692	0.559	0.457	0.547	0.346	0.396
12	0.515	0.270	0.749	0.219	0.738	0.223
13	0.540	0.326	0.659	0.256	0.753	0.124
14	0.664	0.558	0.382	0.529	0.402	0.124
15	0.706	0.692	0.174	0.670	0.255	-0.014
RMR	0.069	0.047		0.035		

Note: Loadings for two- and three-factor solutions are Varimax rotations. Loadings less than 0.4 are greyed.

If these items reflect a single 'course quality' factor, one factor should account for the pattern of correlations among the 15 items. With the exception of Item 1 (course workload), the remaining items show reasonable loadings on a common factor. Items 12 and 13 have rather modest loadings. The root mean square residual (RMR) for this solution is 0.069 and is a

little high for a good solution. In the two- and three-factor solutions, the fit is acceptable (RMR = 0.047 and 0.035 respectively), although Item 1 again fails to load on any of the extracted factors. There is some separation of items 10, 11, 12 and 13, but the separation does not lead to a readily interpretable solution. No items load strongly onto the third factor, so this is not an acceptable solution. It appears that the two-factor solution is the best of the three investigated, but only two items load more strongly onto this factor than the first. Other analyses (see below) shed light on this problem.

Confirmatory factor analysis

The 15 items in the Course SELT instrument were subjected to confirmatory factor analyses (CFA) using LISREL (Joreskog & Sorbom, 2007). A single latent factor (perceived course quality) was postulated to explain the covariances of these items and a single factor model was constructed and then subjected to a refinement process. The refinement process involved removing items that had standardised loadings below 0.40 and for which a unique factor was apparent. The revised model was then subject to a further CFA run. Items with loadings of or below 0.40 were indicative of poor scale fit. Items were not removed simultaneously, however. In the Course SELT instrument, item 1, a course workload item, was removed first due to its very low loading of 0.01. This means the item is not reflective of the model's common factor. Items 12 (absence of discrimination) and 13 (students background) with loadings of 0.43 and 0.45 respectively, were considered modest and they were removed. It is believed that these items have strong similarity in that they measure an important but unique factor. This is discussed below. Item loadings for each CFA run are shown in Table 8.

Table 8: Factor loadings for the Course SELT items after three CFA runs

	CFA runs – items retained		
Item	15	14	12
1	0.01	-	-
2	0.80	0.80	0.81
3	0.82	0.82	0.83
4	0.70	0.70	0.70
5	0.69	0.69	0.69
6	0.61	0.61	0.61
7	0.85	0.85	0.86
8	0.72	0.72	0.72
9	0.70	0.70	0.72
10	0.59	0.59	0.58
11	0.65	0.65	0.63
12	0.43	0.43	-
13	0.45	0.45	-
14	0.62	0.63	0.61
15	0.69	0.69	0.69

Summary of the results of the CFA on the Course SELT single factor model are shown in Table 7.6. It includes the 'absolute fit indices' (Diamantopoulos & Siguaw, 2000, p.87) that assess how well the sample covariances were reproduced by the covariances predicted from the parameter estimates. These indices are the Root Mean Square Error of Approximation (RMSEA), Goodness-of –fit Index (GFI), Adjusted Goodness-of-fit Index (AGFI) and the Root Mean Square Residual (RMR). The Parsimony Goodness-of-fit Index (PGFI), which indicates model complexity, is also included. This set of fit indices in the single factor model with 12 items remaining shows a slight improvement over the other two models. This model yields an acceptable structural model that provides a basis for true measurement. Although the RMSEA, which is generally considered as the most informative of the fit indices (Diamantopoulos & Siguaw, 2000), exceeded the 0.05 threshold indicative of good fit, the GFI, AGFI and RMR all show an indication of a good solution. The RMSEA value equal to

0.064 indicates a reasonable fit (Diamantopoulos & Siguaw, 2000). Hu and Bentler (1999) suggest that values for the RMSEA and RMR below 0.60 and 0.80 respectively are indicative of good fit. They also suggest that good model fit is indicated by values of more than 0.95 for the TLI, CFI and RFI.

Table 9: Summary of CFA results on Course SELT single factor model.

Model	Variables retained	Chi-square	df	GFI	AGFI	TLI	RMR	RMSEA
Single Factor	15	9527.57	90	0.84	0.79	0.949	0.064	0.066
Single Factor	14	8939.86	77	0.84	0.78	0.952	0.066	0.067
Single Factor	12	5702.37	54	0.88	0.82	0.964	0.050	0.064

The threshold value for the GFI and the AGFI is 0.90, and the RMR 0.05 or less. Although GFI value of 0.88 and an AGFI value of 0.82 are less than the threshold value of 0.90, they are close enough to merit a reasonable fitting model. In addition, the TLI is above the minimally acceptable value of 0.90. The model can still be improved. In other words, items in the model could either be removed or correlated to improve the model's fit. However, this would introduce complexities to the analysis of the model.

In its present state the Course SELT questionnaire appears to be fit for purpose.

Rasch analysis

Rasch analyses were undertaken in three stages. The initial analysis included all items and persons. Several items were found not to conform to the requirements of measurement and these items were removed, one at a time. In the sections that follow, in which the Course SELT instrument is evaluated at successively finer levels, information on the initial and final solutions is presented. Although the three levels of analysis are presented separately, interpretation of the results of these analyses is iterative as information gained at one level assists in understanding the results that arise from other levels.

The macro level: the Course SELT scale

The Course SELT scale formed from its constituent items provides a measure of student perception of course quality.

In classical item analysis, Cronbach's Alpha is taken as an indicator of scale consistency, with values over 0.7 taken as indicating acceptable fit. In the Rasch model, two indices are reported, the item reliability index and person reliability index. The former indicates the extent to which the items of the scale cohere to provide a measure of a common construct while the person reliability index, which is normally close to the value of Cronbach's Alpha (Andrich, 1982), indicates the extent to which individuals responses cohere.

Using all 15 items, the item reliability index is 0.95 and the person reliability index is 0.90. With items 1 (workload), 12 (absence of discrimination) and 13 (students' backgrounds) removed, the item reliability index is 0.95 and the person reliability index is 0.91. These values indicate an effective measurement scale. The differences between the indices for the full set of 15 items and the reduced set of 12 items are very small. The item reliability index is at a ceiling level for this scale because of the large number of cases that contribute to it. The decision to remove the three items was based on their individual fit statistics (see below) and not on the overall scale indices.

In addition to scale and person reliability indices, a third characteristic of the instrument is the extent to which it is appropriately targeted for the population of respondents. There is no natural scale for perceived course quality – the construct that is of interest in administering

the Course SELT instrument. By default, the Rasch software being used, Quest (Adams & Khoo, 1999), sets the zero point of the scale at the midpoint of the item difficulties. The interval scale that emerges from the analysis has been rescaled to a person mean of 500 with a standard distribution of 100 points.

The relationship between the distribution of persons and item thresholds is shown in Figure 2. This is a 'cave' plot, including a column graph (stalagmites) for the distribution of person scores and an inverted column graph (stalactites) for the distribution of item thresholds. These thresholds are the reference points that provide measures for individuals.

Figure 2 reveals that the Course SELT instrument is targeted at people who have lower perceptions of course quality than these students do. Many of the item thresholds are located in the lower region of the distribution of student responses and there are few thresholds in the upper region. That is, the instrument would work very well in an institution in which students had lower opinions of course quality than is revealed for this sample of University of Adelaide students.

The person distribution is approximately normal. The mean was set to 500 and the standard deviation to 100 units. A ceiling effect is apparent, as there is a small group of students clustered around a scale score of 950 units. These people responded in the top category to all items in the instrument. While this level of response is reassuring, it is not possible to determine just how favourable their course perceptions are: for some students, they may be even more favourable than the instrument has allowed them to indicate.

The distribution of item thresholds is non-normal. This does not matter; it is common to find almost uniform distributions of item thresholds in many instruments. Ideally, the thresholds should be spread to cover the range of responses observed among students. This is almost the case, but there is a concentration of thresholds at 350 to 400 units on the scale. The mean item threshold value is 438 or 0.6 standard deviations below the person mean. If the scale were perfectly targeted, the mean values for both respondents and items would be 500. Using 'close' simulations for real data sets, Curtis (2004) showed that case means could be up to about 1 standard deviation from the item mean without seriously compromising scale measurement, but that case means should be within about one half a standard deviation for targeting to be good. The targeting for this scale falls within an acceptable range, but the instrument targeting could be improved to bring the means to within 0.5 of a standard deviation and improve the precision of measurement. Items that are more appropriate for students with higher perceptions of course quality would address this issue and would enable differentiation among those students. Such information could contribute more effectively to continuous quality improvement in teaching and learning.

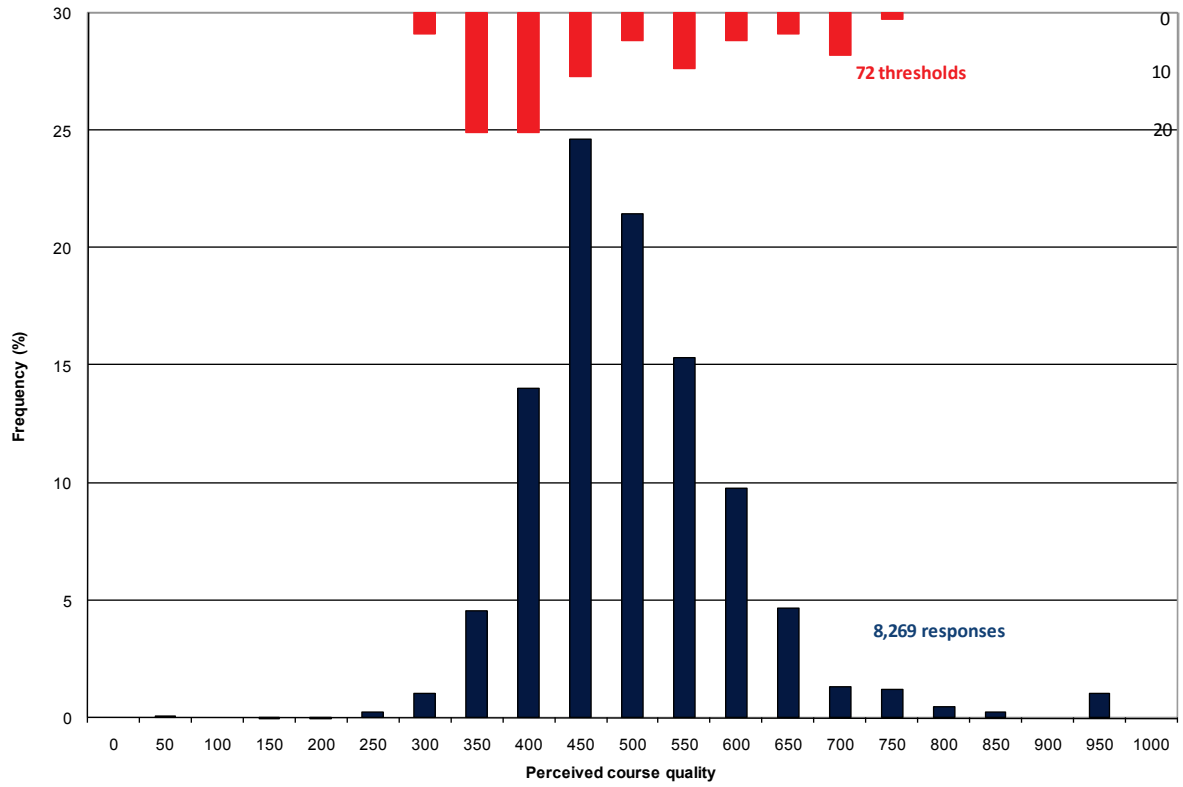

Figure 2: 'Cave' plot of item and person distributions for the Course SELT

In summary, the responses to the Course SELT form a very good 'perceived course quality' scale. There is scope to improve the scale by targeting it more closely to the observed distribution of student perceptions.

The meso level: item and person fit to the measurement model

At the meso level of analysis, we are interested in the fit of items and individuals' responses to the measurement model. Item fit to the Rasch model is judged using a residuals-based indicator, two of which are available. The unweighted mean square fit measure is based directly on residuals between the observed and measurement model predicted responses. This is influenced substantially by outliers, so the information weighted mean square indicator (Infit MS) is most-often reported. The mean square fit values are expected to be about 1.0, with some tolerance either side of this figure. In high stakes testing, it is common to apply quite tight fit criteria, perhaps requiring that Infit values fall between 0.90 and 1.10 units. In attitude survey instruments, Infit values up to about 1.3 are often accepted. Given that responses to these instruments are a low-stakes activity for students, we have taken fit indices of between 0.7 and 1.3 as indicating acceptable fit. The measurement of fit in Rasch analysis is relative, not absolute. Changing the content, be adding or removing other items, changes the fit of individual items in an instrument. Fit indices are not the only information used to make judgments about item acceptability. Because the construct (perceived course quality) being evaluated through the Course SELT is multi-faceted, some qualitative judgment taking into account individual item content, must be exercised.

Of the 15 items in the Course SELT, three showed some misfit. The assessment of fit was evaluated in three stages and the results of these fit analyses are shown in Table 10. In particular, the first item, on workload, showed quite poor fit. This was removed and the remaining items were subject to a second analysis round. In this iteration, two items, item 12 (absence of discrimination) and 13 (students' backgrounds) revealed misfit and were removed. We do not regard them as poor items: they are very good. A close analysis of their

response patterns suggests to us that they may have particular salience for some students. In the absence of any demographic data on respondents, for example gender, home language or home country, we are unable to conduct a detailed analysis of responses by likely groups.

We believe that these items are important, and we support their retention in the instrument. More value could be extracted from them if other data on respondents were available to support their interpretation.

We note that following the removal of items 1, 12 and 13, item 6 (feedback on assignments) falls just outside the fit criteria that we established. This misfit shows the influence of some items on the relative fit of others. Assessment is a facet that is central to course quality, and taking into account the distribution of its thresholds (see below), we decided to retain this item in the analysis.

Table 10: Item fit indices for the Course SELT instrument

	Iteration 1, 15 items		Iteration 2, 14 items		Iteration 3, 12 items	
Item	Information-weighted mean square	Unweighted mean square	Information-weighted mean square	Unweighted mean square	Information-weighted mean square	Unweighted mean square
1	2.06	2.17				
2	0.73	0.72	0.78	0.77	0.79	0.78
3	0.78	0.77	0.82	0.83	0.82	0.82
4	0.90	0.91	0.99	1.02	1.04	1.07
5	0.89	0.88	0.95	0.95	1.00	0.99
6	1.14	1.16	1.25	1.30	1.32	1.38
7	0.68	0.68	0.71	0.71	0.69	0.69
8	0.85	0.85	0.91	0.92	0.94	0.94
9	0.82	0.84	0.88	0.93	0.92	0.96
10	1.01	1.03	1.09	1.15	1.20	1.25
11	0.89	0.88	0.95	0.95	1.06	1.05
12	1.24	1.21	1.34	1.35		
13	1.20	1.20	1.30	1.35		
14	0.93	0.92	1.01	1.01	1.13	1.12
15	0.93	0.93	0.96	0.98	1.01	1.02

In addition to estimates of item fit and the precision of the estimates of item thresholds, the precision of person estimates is also given. Typically, the error of measurement is about 27 units on a scale with a mean of 500 and a standard deviation of 100 units.

Response patterns of individuals have been examined for consistency. If students hold a particular level of perception of course quality, their responses should consistently reflect that level. Some students, however, may endorse a strongly favourable response to one item and a very unfavourable one to the next. This may be a true reflection of their experiences, but occasionally response patterns are seen that suggest careless or even deliberate inconsistent patterns, such as alternating between the most and least favourable options. While in some other survey instruments such patterns appear in almost 20 per cent of cases, the incidence of aberrant responses is less than 10 per cent and serious inconsistencies are seen in fewer than five per cent of cases.

In high stakes testing, such cases may be excluded from analyses. In this situation where student responses may indicate a diversity of experiences with different aspects of a course, there are few responses that could be rejected on person misfit grounds. We find that where responses are inconsistent, the error of measurement is higher. In some later analyses using the Rasch scaled estimates, we use the inverse of the standard error to weight responses. In this way, data from all respondents are used, but those that appear to be more reliable are weighted more heavily.

The micro level: item thresholds

The fundamental idea behind Rasch measurement is that individuals have a perception of course quality. Among students and across courses, we expect this perception to vary continuously from quite low to quite high. We anticipate that students will express this perception through the choices of response options that they make when prompted by the various items in the instrument. Students with low perceptions of course quality are most likely to endorse the less favourable options while students with high opinions of their courses will agree with the more favourable options. A question that we resolve through the Rasch model is 'What level of course quality perception is required before a student is most likely to endorse a more favourable option than the adjacent less favourable one?' That point is the threshold between adjacent response options. Effective instruments have dispersed thresholds that, over the set of items, cover the range of student perceptions. From Figure 2, the effective range of student perceptions is from 250 to 750 units, with a group of students scored at about 950 units. The trait ranges covered by the response options, and separated by the thresholds, for each item are shown in Figure 3.

Very few individuals had course quality perceptions below 250 units. The point of separation between the lowest two response options – the lowest threshold – is at about 300 units. This threshold differentiates very few students, and so does not make a particularly effective contribution to the measurement provided by the instrument for this population. In effect, the lowest response option attracts relatively few responses (fewer than 2%, see Table 5).

The mean of the scale was set at 500 units and with an approximately normal distribution, half of the students have perception of course quality scores above this level. However, typically there are only two thresholds for each item above this level. The instrument is targeted at below average responses. It would work well in institutions where students held less favourable views of course quality than these students have. With thresholds clustered below the mean, the regions of the scale covered by these lower response options are relatively narrow and each operates to measure a modest proportion of respondents well. For item 14, for example, the bandwidth of the third response option is extremely narrow and the adjacent thresholds are not statistically distinct.

Effective measurement requires dispersed thresholds. Maintaining some of the items as they are, but making some other items more difficult to endorse at high levels, perhaps by changing the item prompt, would move some of the response regions to higher levels on the scale and provide better measures of the many students who have favourable perceptions of the courses they take.

As an example of what could be done, the current item "I am motivated to learn in this course" could be altered to "I am strongly motivated to learn in this course." Only those students with very high perceptions of course quality would select the higher response categories of this item. This change is not being recommended: it illustrates that a simple change could be made to alter the response characteristics of the scale. Any such changes should follow a detailed construct analysis of the items and should be pilot tested before they are incorporated into the instrument.

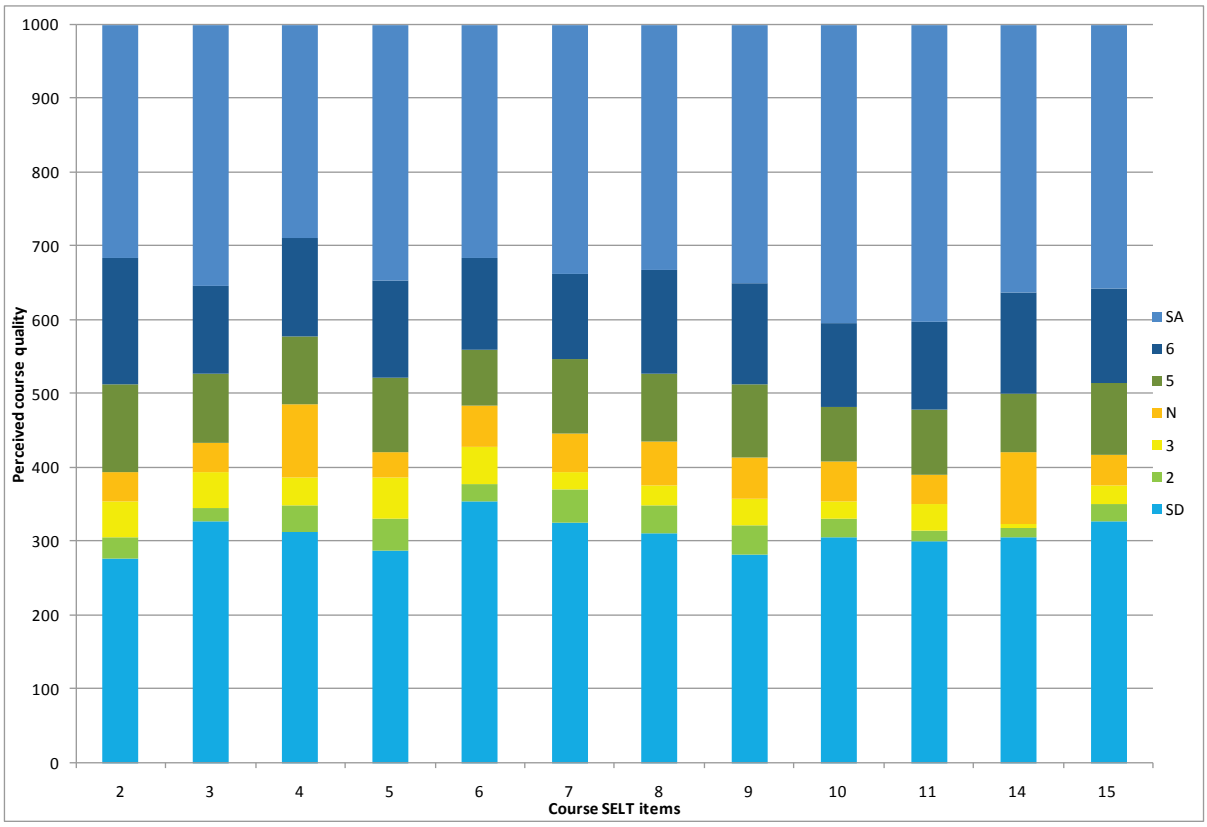

Figure 3: Response distribution map for the Course SELT

Summary of Rasch analyses of the Course SELT data

Rasch analysis has been used to test the coherence of the Course SELT instrument.

Overall, the scale has good measurement properties and provides useful information about course quality. The information cold be used to monitor perceptions of course quality over time.

The scale could be improved. It is slightly 'off-target' in that many students' perceptions of their courses are more favourable than might have been anticipated in the construction of this instrument.

For some students, an instrument-imposed ceiling is evident. Adding items that are a more rigorous test of course quality or modifying some existing items could retarget the instrument.

The Rasch analysis has been used to estimate students' perceptions of course quality on an interval scale. In addition to estimating students' scores, the scores are estimated with a margin of error and this can be used to decide whether the estimates are sufficiently robust to support the judgements that are made using these data.

Teacher SELT

The Teacher SELT comprises seven items, each with a seven-point Likert response format. The text of the items is shown is Table 2. The extreme and mid-point responses for the first item are labelled 'very poor', 'outstanding' and 'reasonable' respectively. For subsequent items, the corresponding labels are 'strongly disagree', 'strongly agree' and 'undecided'. For all items, the intermediate options are numbered and there is a 'not applicable' option.

Descriptive summary

Frequencies of the various response options to the Teacher SELT items are shown in Table 11. As was seen in the Course SELT results, few students choose the lowest three options and option 6, the second highest category, is the modal response category for all seven items. In this table, mean response values are presented.

Table 11: Response frequencies for the Teacher SELT items

	Response option labels									
Item	Very poor	2	3	Reason-able	5	6	Out-standing	N/A	Missing	Mean
1	171	275	569	2075	4204	6789	3593	54	175	5.52
2	134	202	499	1346	3523	7018	5103	56	24	5.77
3	170	283	674	2396	4444	5796	3920	173	49	5.47
4	188	277	687	1860	3956	5606	5210	88	33	5.63
5	220	419	1034	2861	4329	4781	3853	373	35	5.31
6	512	587	1145	2623	4200	4866	3849	86	37	5.22
7	327	457	812	1815	3565	5718	5098	56	57	5.55

Note: Data from 17,905 Teacher SELT questionnaires were available.
The text for these items is shown in Table 2.

Cronbach's alpha for the set of items was found to be 0.933 (see Table 12). This is quite high, especially given that there are only seven items, and suggests that the set of items cohere to form a strong scale. The item-total correlations are all quite high and there is no case for removing any of the items.

Table 12: Results of scale reliability analysis for the Teacher SELT

Item no.	Scale mean if Item deleted	Scale variance if Item deleted	Corrected Item-total correlation	Cronbach's Alpha if Item deleted
1	32.99	45.15	0.85	0.918
2	32.75	47.74	0.71	0.930
3	33.04	45.78	0.77	0.924
4	32.89	44.65	0.82	0.920
5	33.20	45.09	0.74	0.928
6	33.30	42.11	0.83	0.919
7	32.97	43.92	0.79	0.923
Scale alpha				0.933

Exploratory analysis

An exploratory factor analysis was undertaken on the Teacher SELT data using Mplus (Muthen & Muthen, 2006) and the results of this analysis are shown in Table 13. Only a one-factor solution was tried as the factor loadings were high and the fit of that single factor was good, indicated by an RMR of 0.042.

Table 13: Factor loadings for the Teacher SELT

Item	Loading
1	0.910
2	0.773
3	0.832
4	0.881
5	0.795
6	0.886
7	0.844
RMR	0.042

Confirmatory factor analysis

The seven items in the Teacher SELT instrument were subjected to confirmatory factor analysis (CFA) using LISREL (Joreskog & Sorbom, 2007). A single factor model was constructed and the refinement process was not carried out on the Teacher SELT because each item satisfactorily loaded onto a single factor and the set of items showed generally reasonable fit to a single latent perceived quality of teaching factor. Item loadings for the CFA run are shown in Table 12.

Table 14: Factor loadings for the Teacher SELT items

Item	Loading
1	0.91
2	0.78
3	0.82
4	0.87
5	0.79
6	0.90
7	0.86

With each of the seven items loading at least 0.78, it appears that they form a coherent scale. With the exception of the RMR, the fit is not clearly particularly good, as shown in Table 13 with RMSEA showing a value that is quite high (0.120) indicating a poor fit. Teaching quality is a complex construct with many facets. The questions in the Teacher SELT assess some of those facets, and those that are assessed, by not completely representing the construct, necessarily provide pieces of the puzzle that do fit together as neatly as if all the pieces were in place.

Table 15: Summary of CFA results on Teacher SELT single factor model

Model	Variables retained	Chi-square	df	GFI	AGFI	TLI	RMR	RMSEA
Single Factor	7	7271.5	14	0.878	0.757	0.974	0.037	0.120

To indicate good fit, a model requires its GFI and AGFI to have a value of at least 0.90. The Teacher SELT single factor model has a GFI that is slightly below 0.90 and an AGFI that is significantly lower than 0.90. The PGFI, which takes account of a model's complexity, is very low. The TLI value, at 0.974, is above the minimally acceptable value of 0.90. Teaching is a complex and multifaceted construct. It is unlikely that it can be captured satisfactorily in seven items and a case can be made for adding items to capture additional dimensions of this activity. **However, caution should be strongly considered when adding items so that the instrument does not become onerous for students**. In the Teacher SELT instrument's current state, the RMR is below 0.05, which indicates a good fitting model, but other fit indices suggest less than adequate fit. Generally, based on the CFA results, the instrument appears to serve its intended purpose, but it could be improved.

Rasch analysis

An initial analysis included all items and persons. All items conformed to the requirements of measurement so no item deletions were required.

As was the case for the Course SELT, the Teacher SELT instrument is evaluated at successively finer levels. Although the three levels of analysis are presented separately, interpretation of the results of these analyses is iterative as information gained at one level assists in understanding the results that arise from other levels.

The macro level: the Teacher SELT scale

The Teacher SELT scale formed from its constituent items provides a measure of student perception of teaching quality.

Using all seven items, the item reliability index is 0.97 and the person reliability index is 0.89. The item reliability index is slightly flattering because of the large number of cases analysed, but it is very good. The case reliability index is very good, considering that there are only seven items.

Targeting is a problem with the Teacher SELT. The relationship between the distribution of persons and item thresholds is shown in Figure 4.

The person distribution is approximately normal over much of the range, but a strong ceiling effect is apparent with a substantial group of students (10% of the sample) clustered around a scale score of 700 units. The range of scaled scores for the Teacher SELT is truncated compared with the Course SELT, reflecting the smaller number of items. These people responded in the top category to all items in the instrument. It is reassuring that so many students have very favourable perceptions of their academic teachers, but it would be more useful to discriminate among them. The perceptions of the most satisfied students would provide useful information for continuous quality improvement in teaching.

The mean item threshold value is 400 or 1.0 standard deviation below the person mean. If the scale were perfectly targeted, the mean values for both respondents and items would be 500. The targeting of the instrument to the observed sample is at the outer margin of acceptability. Items that are more appropriate for students with higher perceptions of teaching quality would address this issue and would enable differentiation among those students. Such information would contribute more effectively to continuous quality improvement in teaching and learning.

In summary, the responses to the Teacher SELT suggest that the items all cohere to form a common measure of ‘perceived teaching quality.’ However, the instrument is not well targeted. Some existing item could be amended and, given that the instrument has only seven items, some new ones could be added without imposing an excessive burden on students.

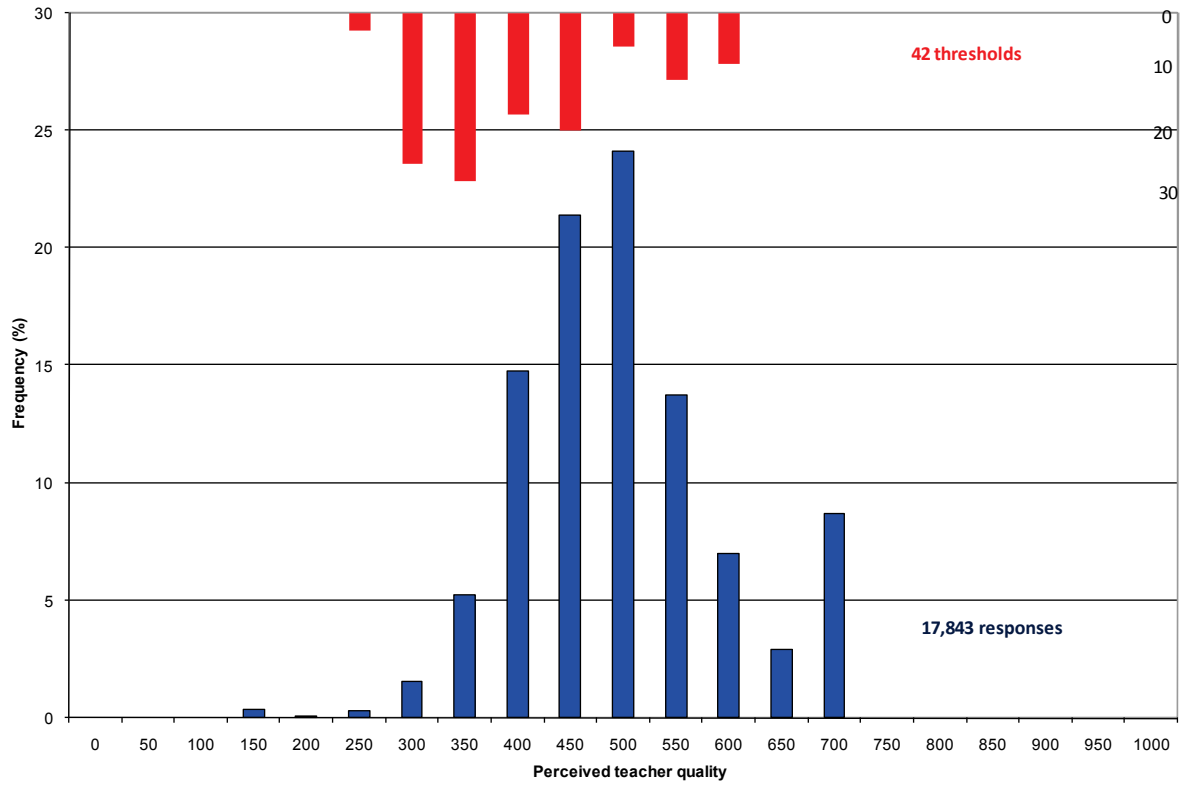

Figure 4: 'Cave' plot of item and person distributions for the Teacher SELT

The meso level: item and person fit to the measurement model

At the meso level of analysis, we are interested in the fit of items and individuals' responses to the measurement model. Given low-stakes nature of the instrument for students, we have taken fit indices of between 0.7 and 1.3 as indicating acceptable fit. None of the seven Teacher SELT items showed misfit (see Table 16).

Table 16: Item fit indices for the Teacher SELT

Item	Information-weighted mean square	Unweighted mean square
1	0.70	0.72
2	1.24	1.23
3	1.04	1.04
4	0.80	0.80
5	1.23	1.23
6	0.84	0.84
7	1.01	1.00

In addition to estimates of item fit and the precision of the estimates of item thresholds, the precision of person estimates is also given. Typically, the error of measurement is about 25 units on a scale with a mean of 500 and a standard deviation of 100 units. Those students whose estimated perception of teaching scores are high (around 700 units), and at some distance from the cluster of item thresholds, have higher measurement errors in their estimates.

The incidence of aberrant responses is less than 10 per cent and serious inconsistencies are seen in fewer than two per cent of cases. This is quite good. It suggests that the great majority of students take the teacher SELT seriously. It would be unreasonable to

expect all students to do this. As was the case for the Course SELT, we use the inverse of standard errors to weight responses in later analyses.

The micro level: item thresholds

From Figure 4, the effective range of student perceptions is from 300 to 700 units, with a group of students clustered at a score of about 700 units. The trait ranges covered by the response options, and separated by the thresholds for each item, are shown in Figure 5.

Very few individuals had teaching quality perceptions below 300 units. The point of separation between the lowest two response options – the lowest threshold – is at about just less than 300 units. This threshold differentiates very few students, and so does not make a particularly effective contribution to the measurement provided by the instrument for this population. In effect, the lowest response option attracts relatively few responses (between 1% and 3%, see Table 11).

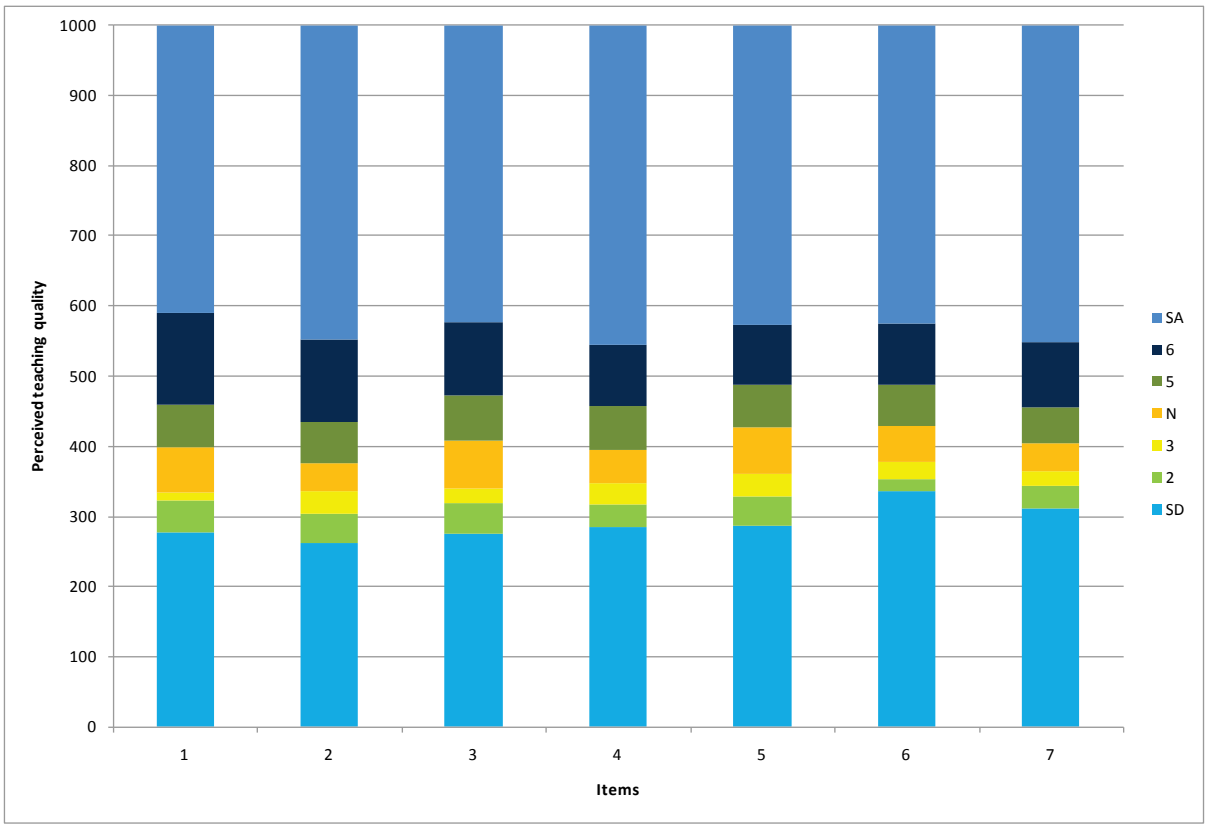

Figure 5: Response distribution map for the Teacher SELT

The mean of the scale was set at 500 units and with an approximately normal distribution, half of the students have perception of course quality scores above this level. However, typically there are only two thresholds for each item above this level. The instrument is targeted at below average responses. With thresholds clustered below the mean, the regions of the scale covered by these lower response options are relatively narrow and each operates to measure a modest proportion of respondents well. For most items, the bandwidth of the third response option is quite narrow, especially for the first item, and for this item the adjacent thresholds are not statistically distinct.

Effective measurement requires dispersed thresholds. Maintaining some of the items as they are, but making some other items more difficult to endorse at high levels, perhaps by changing the item prompt, would move some of the response regions to higher levels on the scale and provide better measures of the many students who have favourable perceptions of the courses they take.

Summary of Rasch analyses of the Teacher SELT

Rasch analysis has been used to test the coherence of the Teacher SELT instrument.

Overall, the items cohere to measure a common perceived quality of teaching construct.

The scale could be improved. It is substantially 'off-target' in that many students' perceptions of their courses are more favourable than might have been anticipated in the construction of this instrument.

For about 10 per cent of students, an instrument-imposed ceiling is evident. Adding items that are a more rigorous test of teaching quality or modifying some existing items could retarget the instrument.

In addition to testing scale coherence, the Rasch method has been used to generate individual scores with known measurement errors. The fit of individual responses to the requirements of measurement has shown that only a very small proportion of students have provided aberrant responses.

4 Supplementary Analyses

The Rasch analyses have generated interval scaled scores for individuals. These scores can be used in other analyses that require at least interval data. Examples of the sorts of analyses that can be undertaken are presented as indicators of possible analyses.

What is the relationship between workload and perceived course quality?

In the Course SELT, we found that the first item, about workloads, was not related to the other items in the scale. In calculating scores for the perceived course quality construct, this item was removed. We are interested in the perceived course quality associated with each level of the response options to this item. This relationship is shown in Figure 6. In this analysis, values of perceived course quality were weighted using the inverse of the standard error of measurement for each individual. Because the scale is not as well targeted as we might hope, higher levels of this trait tend to be weighted down more than lower values, so the mean of the weighted score is about 480 rather than 500 units for the unweighted score. The figure shows the mean value of the weighted score for each level of the workload item. The vertical bars show the confidence intervals for the mean. The longer error bars reflect small numbers of students responding in those categories. Most students think the workload is close to about right.

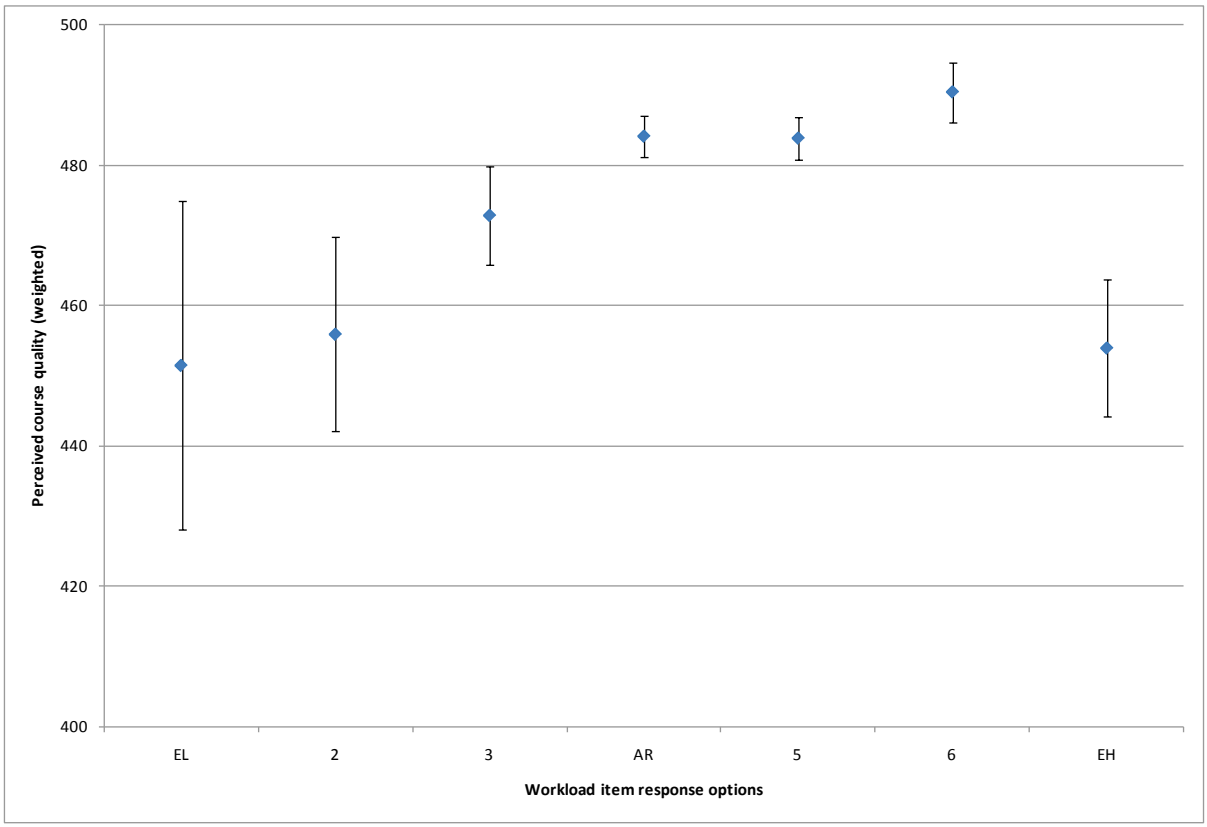

Figure 6: Relationship between perceived course quality and perceptions of course workload

The response options available to students were ‘extremely light’, ‘about right’ and ‘extremely heavy’ with intervening levels indicated by numeric labels. Overall, the

relationship is strongly curvilinear, and this explains the very low correlation between this item and others in the Course SELT scale.

Several interpretations of this relationship are apparent. One possibility is that most courses have about the right workload, in a few it is too low, and in some is very heavy. In those with extremely heavy workloads, students find the workload detracts from other aspects of the course quality. That is, this explanation infers causal link from workload to perceived quality. An alternative view is that workload equates to challenge, and that as the demand of courses increases, students find the level of challenge equates to other facets of course quality. However, when workloads become unreasonable, the relationship between challenge and perceived quality is fractured. Other explanations might posit that when students are unhappy with a course on other grounds – those revealed by the other items in the Course SELT scale – they express their frustration by saying that they workload was either too light or too heavy. The inference of causality can run in either direction. Neither explanation is supported by the data: the inference of causality is imputed by the observer.

In summary, the relationship between workload and other facets of courses is curvilinear and this explains the lack of a net correlation between this and other items. The non-linear nature of this relationship invites interpretation.

By how much do perceptions of course quality vary between courses?

We observe some variation between students in the perception of course quality and we expect to see some of this variation reflected between courses. The data used in this study was 'confidentialised' and we know little about the various courses from which the data were collected. We do know that the data come from 17 disciplines (representing 14 schools and 5 faculties) over three calendar years and we know how many students were enrolled in these courses. Collectively, the data analysed represent 47 different course presentations. The mean value of students' perceptions of course quality for each of these course presentations is shown in Figure 7. The weighted values of perceived course quality are plotted, the weighting reflecting the reliability of the individual measures. The mean weighted value of perceived course quality is 480 (compared with 500 units for the unweighted variable).

We do not know what the disciplines are, nor whether they are strongly qualitative or quantitative in orientation, nor what the gender or age distributions of students in these courses were. These are factors that are known to influence student judgments of course quality (Curtis & Keeves, 2000; Meyler, 1997).

The mean value of perceived course quality is indicated in the figure with a labelled horizontal line. The mean value lies at the centre of a band that reflects courses that, within the constraints of sampling and measurement error, are not significantly different from the mean. On either of this central band lie regions that represent course that are significantly above and below the mean value. It is apparent that there are very few courses below the lower limit of the average band. There are more courses above this average band, and two or three courses that have been rated especially highly by students.

The length of the standard error bars in the plot reflects the numbers of students in those courses. It is apparent that the courses rated most favourably by students tend to be relatively small and that those rated more harshly have larger enrolments. This led us to investigate the influence of course size on student perception of course quality. We found a statistically significant but very weak effect. Large course enrolments are associated with lower levels of satisfaction, but it accounts for a very small proportion of the variance in student perceptions (R^2=2%). It explains very little of the differences that are observed between courses.

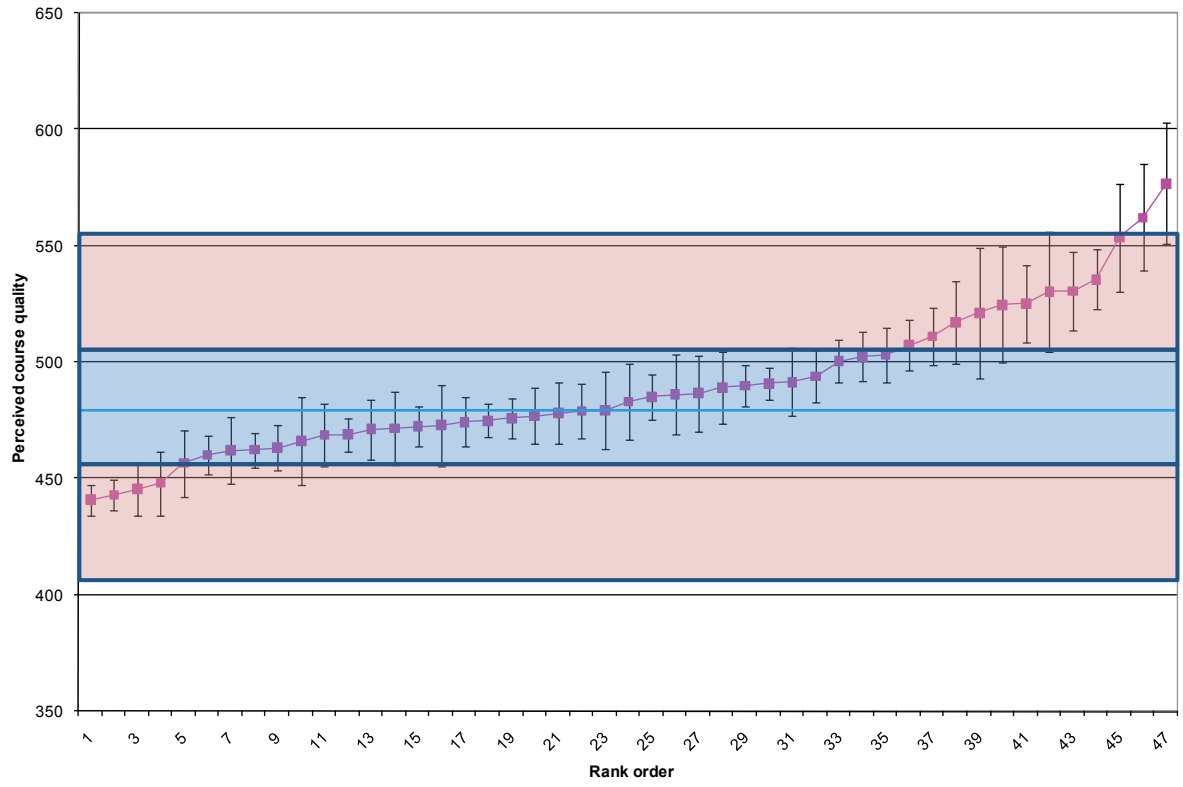

Figure 7: Mean values and confidence intervals of perceived course quality by course

By how much do perceptions of teaching quality vary between teachers?

We compared the perceptions that students have of their teachers in the same way that we compared their course perceptions. Weighted individual teaching quality perception scores were used, and because of the mistargeting of the instrument, the more favourable perceptions had greater errors of measurement than the less favourable ones, and were therefore weighted down. The mean of the weighted 'perception of teaching quality' variable is 475, compared with 500 for the unweighted value.

The Teacher SELT data were collected from 117 administrations of the instrument over three years. Some teachers may be represented several times in different courses and in different years in the data set. Because the data were de-identified, it is not possible to know which SELT administrations are associated with which teachers, so all administrations of the instrument are treated as representing separate teaching episodes. The mean 'perceived teaching quality' score for each administration of the SELT are shown in Figure 8. In that figure, 95% confidence intervals are shown for each instance in which the Teacher SELT was administered. However, in addition to sampling error, there is also some measurement error. Although it is modest, it is taken into account in setting the performance bands about the mean value.

Individual students vary in their perceptions of the teachers they experience. However, there are substantial differences between teachers. Over half of the Teacher SELT instances lie within the average perceived teaching quality band. About 20 per cent of cases lie below this band and 20 per cent above it. Three cases lie below the less-than-average band and six are located above the higher-than-average band.

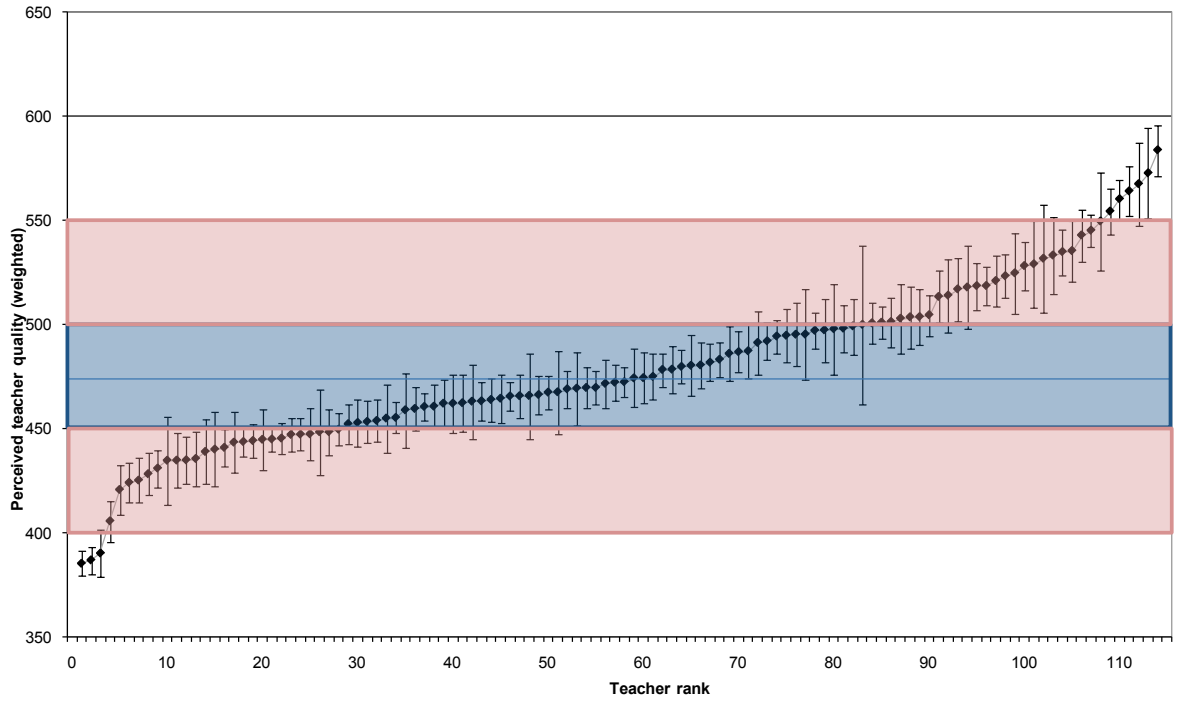

Figure 8: Mean values and confidence intervals of perceived teaching quality by teacher

If more information about the teaching practices and contexts were available, additional analyses that use the features of courses and teachers could be undertaken. Such analyses may contribute to the University's continuous quality improvement processes for teaching.

5 Conclusions

In this section, we summarise the key findings of our investigation, we present some conclusions that arise from the analyses presented above, and we draw attention to some possible implications that arise from this study.

Key findings

Both the Course and Teacher SELT instruments have quite reasonable measurement properties. Decisions based on them in their present form are soundly based.

The instruments are not ideal. In both, there is scope for improvement. The Course SELT includes an item on workload. This is not related to other items in the instrument. Its inclusion does not compromise the Course SELT, but it does not contribute to the measurement of perceived course quality.

Two items, one about the absence of discrimination and the other about considering students backgrounds, are related to each other. They do not show a strong relationship with other items in the instrument, but they may have particular salience for sub-groups of the student body. They are worthy items and would be more meaningful if additional information were available about the students who respond, perhaps information on gender or country of birth.

The Teacher SELT is a brief but informative instrument. All seven items cohere well and contribute to the measurement of students' perceptions of the quality of teaching. The construct is not measured comprehensively. We believe that the inclusion of some additional items could improve this instrument.

Both instruments, especially the Teacher SELT, could be better targeted to the high esteem that students have for the quality of courses and teaching they experience at the University of Adelaide. The mistargeting of the instruments compromises the response categories. That is, although there are seven response options, they do not all operate effectively. It is very likely that having fewer response categories would yield similar information to that obtained from the current instrument.

We find that relatively few students (<5%) provide inconsistent responses to these instruments. This suggests that students take the evaluation of courses and teaching seriously and their responses make a substantial contribution to the University's quality improvement processes.

The Course SELT

Do the items of Course SELT function as expected?

Most items in the Course SELT function well. While we make some suggestions that we believe could improve the Course SELT, we believe the existing instrument has provided a sound basis for assessing course quality. We make several suggestions that could improve the value of the instrument for the University's quality improvement processes.

Workload: We find that one item, seeking students' perceptions of workload, did not contribute meaningfully to the Course SELT scale. This item may provide useful information if responses to it are considered separately from responses to other items. The interpretation

of responses is not simple. In isolation, a course coordinator may make a judgement about most students' perceptions of workload as being either too light, too heavy or about right. In general, many students who had relatively high perceptions of course quality found the workload about right. But those who had low perceptions of quality may have rated the workload as either being too heavy or too light. This is unhelpful to coordinators who are trying to improve courses.

Students expect and deserve to be challenged. Too little challenge means that there is a reduced opportunity for learning, but too much challenge, perhaps because of an over-crowded curriculum, means that students have little opportunity to internalise elements of the course before the next set of tasks are set. In either case, learning is compromised.

There is a case for reviewing the workload item. The issue of challenge could be addressed directly. A possible item might be 'The concepts presented in the course were new and challenging'. This would not address the possibility of overload. A separate item would be required to detect this. We are not recommending this change, but we do suggest that the matter be investigated. How workloads are monitored in instruments used elsewhere is one approach. A second is to convene focus groups of academic teachers and students to explore the issue and to develop, and then trial, possible alternative items.

Inclusivity: We find that two items, both dealing with aspects of inclusive practices, did not fit well with the remaining items of the Course SELT scale. We believe that these items are important and do not suggest their removal. The measurement problem arising from these items is a result of particularly low frequency responses to some of the response options for these items: That is, their response patterns were different from those observed for the remaining items. We suspect that these items will be much more informative is some demographic information is available with the SELT. For example, those people who suggest that the course was not inclusive could represent important minority-group students. This could be based on gender or country of birth. Asking students for such information may lead to concerns about confidentiality and to reduced response rates, especially in courses with small enrolments. However, the low level of aberrant response patterns suggests that students provide considered responses and that they treat the forms seriously. We doubt that the provision of this information would compromise response rates.

How precisely does the Course SELT reflect students' learning experiences?

We find that 12 of the 15 Course SELT items cohere well to measure a 'perceived course quality' construct. Indices that reflect the measurement of this construct suggest that the measurement is very good. (The item reliability index is 0.95 and the case reliability index is 0.91).

The instrument is reasonably well targeted, although this could be improved. There is no need to seek to alter the targeting as a primary objective of any revision of the instrument, but if other changes are made, they should be designed to improve the targeting.

The precision of measurement appears to be quite acceptable. With perceived course quality measured on a scale with a mean of 500 and a standard deviation of 100, a typical standard error is about 25 units. For people with very high perceptions of course quality, the measurement error is higher than this, a situation that would be improved through better targeting.

Not all courses are perceived to be of equal quality. If we assume that measurement errors have a mean of zero, then it is possible to aggregate individual student perceptions to the course level. Any measurement error in the estimated perceived quality of a course is then much smaller than the errors at the individual level. It is apparent that there are quite substantial and significant differences in the perceived quality of different courses. Many courses are close the mean perceived course quality for all courses included in the data set.

But some courses are perceived to be much better than average and some rather worse than average.

Is there evidence of aberrant patterns in responses to the Course SELT?

There is some evidence of aberrant response patterns to the Course SELT. Such responses, typically selecting a favourable response to some prompts and unfavourable ones to other items, may reflect the experiences of individual students. However, the vast majority of students do not find such inconsistencies in the delivery of the courses. The proportion of students showing these inconsistent responses is quite small at fewer than five per cent of the sample. This is very encouraging, suggesting that students take the SELTs seriously.

The Teacher SELT

Do the Teacher SELT items function as expected?

The Teacher SELT, comprising only seven items, works quite well. All items fit a scale of perceived teaching quality. The scale does not meet the rigorous demands of the congeneric model assessed through CFA. However, quality teaching is a multi-faceted activity, and these facets are represented in the items. If improving the psychometric properties of the scale is a desired objective, the addition of items that tap into facets that are represented by existing items would improve the scale's fit indices.

The CFA factor loadings are reasonably strong, and the Rasch fit indices for the individual items indicate that the items do cohere to measure a common construct.

How precisely does the Teacher SELT reflect students' experiences of teaching?

The Rasch scale indices suggest that the scale provides very good measurement, with an item reliability index of 0.97 and a case reliability index of 0.89.

The Teacher SELT scale is not well targeted. Clearly, students' responses indicate that they are very satisfied with the quality of the teaching that they experience. While this is very reassuring, in order to support continuous quality improvement, it is desirable that the set of items reflect observed perceptions. Given that the instrument includes only seven core items, there is scope to add items that are more difficult to endorse. This will increase the mean difficulty of the set of items to match students' perceptions more closely. A modified scale that anticipates high perceptions of teaching quality will discriminate better among those students who have the most favourable views of teaching quality and, in turn, will enable discrimination among the better teachers.

Typical standard errors of measurement are similar to those found for the Course SELT, at about 25 units on a scale with a mean of 500 and a standard deviation of 100 units. Given that there are only seven items, this is a good outcome. When individual scaled perception scores are aggregated for each teacher, the standard errors of measurement at the teacher level are quite small.

Is there evidence of aberrant patterns in responses to the Teacher SELT?

As was the case for the Course SELT, there evidence that a very small proportion of students (less than 5%) show substantially inconsistent responses. As is the case for the Course SELT, students appear to provide thoughtful responses to the Teacher SELT.

Implications and recommendations

We find that the current Course and Teacher SELT questionnaires are fit for purpose. We also find, however, that there is room for improvement in both. If the instruments are changed, the method used to report results, average SELT scores, would not preserve their

meaning. For example, the Teacher SELT does not adequately tap the full range of student perceptions. If items were added to the instrument, it would be sensible to add items that were more difficult to endorse very favourably. This would enable the cluster of students for whom there is a ceiling effect to be differentiated and their views could inform quality improvement processes in the University. A teacher who continued to operate at their current level of perceived teaching performance would find the average response score to be slightly lower than it is now. The continuity of raw scores over time would be broken.

The solution to this situation is to calibrate the current forms of the two instruments using current responses. Those items that continue unchanged into the future will carry with them their currently calibrated values. Courses and teachers assessed using the revised versions of the questionnaires will have scaled scores that are comparable over time. Any change will thus be meaningfully captured. Further, because measures are estimated with known measurement error, the significance of any change can be estimated.

The use of a measurement approach to the assessment of perceived course and teacher quality will also create new possibilities. The questionnaires are necessarily short and cannot measure comprehensively the two constructs. However, banks of appropriate items could be developed. If the University wanted to target a particular aspect of quality improvement, e.g. assessment, a set of assessment items could be used either across the university or within a faculty or school. Again, the measurement approach would ensure that scaled scores are comparable over time. Once that facet of course or teaching quality had been addressed, a sub-set of items focusing on another facet of teaching could be substituted. In this way, and over time, a range of facets of course and teaching quality could be addressed leading to quality improvement without making the questionnaires onerous for students.

The use of a measurement approach to the evaluation of course and teaching quality generates estimates of perceived course at teaching quality at the individual student level. These measures of quality can then be aggregated at the course and teacher levels to provide information about particular courses and teachers. The estimates, having known measurement errors, can be the basis of fair and valid comparisons between courses and teachers and over time. We have shown how this can be done. Much more can be done with quality measures. The literature on student evaluation of teaching has revealed that students' judgments are influenced by gender, age and country of origin. Certain characteristics of courses, e.g. class size, whether it is compulsory or optional and whether it is quantitative, also influence students' responses. Generating fair and unbiased estimates of course and teaching quality depends on factoring these influences out of judgments. Separating individual and course level influence requires multilevel modelling. Having quality measures is a necessary first step in moving towards this outcome.

We suggest that the University consider adopting a measurement-based approach to the analysis of student evaluations of courses and teaching. The improvement of the current instruments should be an immediate goal. Once achieved, several new possibilities are created. These include the fair and reliable estimation of course and teacher effects and may assist on-going quality improvement processes.

6 References

Adams, R. J., & Khoo, S. T. (1999). Quest: the interactive test analysis system (Version for PISA) [Rasch analysis software]. Melbourne: Australian Council for Educational Research.

Andrich, D. (1978). A rating formulation for ordered response categories. *Psychometrika, 43*(4), 561-573.

Andrich, D. (1982). An index of person separation in latent trait theory, the traditional KR-20 index, and the Guttman scale response pattern. *Educational Research and Perspectives, 9*(1), 95-104.

Arbuckle, J. L. (2007). AMOS (Version 15) [CFA and SEM analysis program]. Chicago, IL: SPSS Inc.

Byrne, B. M. (2001). *Structural equation modeling with AMOS. Basic concepts, applications, and programming*. Mahwah, NJ: Lawrence Erlbaum Associates.

Curtis, D. D. (2004). Comparing classical and contemporary analyses and Rasch measurement. In S. Alagumalai, D. D. Curtis & N. Hungi (Eds.), *Applied Rasch measurement: A book of exemplars* (pp. 181-197). Dordrecht: Springer - Kluwer Academic Publishers.

Curtis, D. D. (2004). Person misfit in attitude surveys: Influences, impacts and implications. *International Education Journal, 5*(2), 125-143.

Curtis, D. D., & Boman, P. (2004). *The identification of misfitting response patterns to, and their influences on the calibration of, attitude survey instruments.* Paper presented at the 12th International Objective Measurement Workshop, Cairns, QLD.

Curtis, D. D., & Boman, P. (2007). X-ray your data with Rasch. *International Education Journal, 8*(2), 249-259.

Curtis, D. D., & Keeves, J. P. (2000). The Course Experience Questionnaire as an Institutional Performance Indicator. *International Education Journal, 1*(2), 73-82.

Diamantopoulos, A., & Siguaw, J. A. (2000). *Introducing LISREL. A guide for the uninitiated*. Thousand Oaks, CA: Sage Publications.

Harwell, M. R., & Gatti, G. G. (2001). Rescaling ordinal data to interval data in educational research. *Review of Educational Research, 71*(1), 105-131.

Hu, L.-T., & Bentler, P. M. (1999). Cuttoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling, 6*(1), 1-55.

Johnson, T. (1997). *The 1996 Course Experience Questionnaire: a report prepared for the Graduate Careers Council of Australia*. Parkville: Graduate Careers Council of Australia.

Joreskog, K. G., & Sorbom, D. (2007). LiSRel for Windows (Version 8.80) [Statistical analysis software]. Chicago: Scientific Software International.

Keeves, J. P., & Masters, G. N. (1999). Issues in educational measurement. In G. N. Masters & J. P. Keeves (Eds.), *Advances in measurement in educational research and assessment* (pp. 268-281). Amsterdam: Pergamon.

Masters, G. N. (1982). A Rasch model for partial credit scoring. *Psychometrika, 47*(2), 149-174.

Meyler, M. (1997, 8 -11 July). *What do SET surveys really measure? And why does it matter?* Paper presented at the HERDSA'97 Advancing International Perspectives conference, Adelaide.

Michell, J. (1997). Quantitative science and the definition of measurement in psychology. *British Journal of Psychology, 88*, 355-383.

Muthen, L. K., & Muthen, B. O. (2006). MPlus (Version 4.0) [Statistical analysis with latent variables]. Los Angeles, CA: Muthen & Muthen.

Rasch, G. (1960, 1980). *Probabalistic models for some intelligence and attainment tests*. Copenhagen, Chicago: Danmarks Pegagogiske Institut, University of Chicago Press.

University of Adelaide. (2007). Student Experience of Learning and Teaching (SELT) Policy. http://www.adelaide.edu.au/policies/?dsn=policy.version;field=data;id=11983;m=view

This report was prepared as an outcome of:

The University of Adelaide
2008 Learning and Teaching Implementation Grant

Professor Geoffrey Crisp, Centre for Learning and Professional Development; Dr David D Curtis, A/P Sivakumar Alagumalai, and Dr Darmawan I Gusti Ngurah, School of Education; Dr Steven Barrett, School of Economics

Proposal to Investigate the Validity, Reliability and Measurement Properties of Teacher and Course SELT Questionnaires