Paper Presented at the Joint AARE/
NZARE Conference, Auckland, 2003
Australian Association
for Research in Education

PhD Students' Evaluations of Research Supervision: Issues, Complexities and Challenges in a Nationwide Australian Experiment in Benchmarking Universities

Herbert W. Marsh and Andrew Martin
Self-concept Enhancement and Learning Facilitation Research Centre
University of Western Sydney, Australia

Kenneth J. Rowe
Australian Council of Educational Research
Australia

Paper presented at NZARE AARE, Auckland, New Zealand November 2003
MAR03801

Paper Presented at the Joint AARE/
NZARE Conference, Auckland, 2003
Australian Association
for Research in Education

PhD Students' Evaluations of Research Supervision: Issues, Complexities and Challenges in a Nationwide Australian Experiment in Benchmarking Universities

Herbert W. Marsh and Andrew Martin
Self-concept Enhancement and Learning Facilitation Research Centre
University of Western Sydney, Australia

Kenneth J. Rowe
Australian Council of Educational Research
Australia

Acknowledgments: This article is based on an earlier paper commissioned and distributed by the Australian Council of Deans and Directors of Graduate Students and presented to a meeting of Australian Pro Vice Chancellors of Research and, subsequently, at the Postgraduate Research Experience Questionnaire (PREQ) Seminar that was organized by the Graduate Careers Council of Australia in association with the Australian Department of Education, Training, and Youth Affairs. We would like to thank the Australian Council for Educational Research and Bruce Guthrie of the Graduate Careers Council of Australia who provided us with draft reports and the data used in this article. In addition we would like to acknowledge the Australian Council of Educational Research for their work in assembling and preparing data files that facilitated our research and for earlier analyses of these data conducted by Gerry Elsworth, Adrian Harvey-Beavis, and Sue Fullarton. We would also like to thank Professor Kit-Tai Hau, Kate Johnson, Wijesir Jayasinghe and members of the Australian Council of Deans and Directors of Graduate Studies for helpful comments on earlier versions of this draft. We emphasize, however, that our interpretations of the results are our responsibility and may not be in agreement with those of persons and organizations who we have acknowledged. This research was funded in part through a grant to the first author from the Australian Research Council. Correspondence in relation to this article should be sent to Requests for further information about this investigation should be sent to Professor Herbert W. Marsh, Director, SELF Research Centre, University of Western Sydney, Bankstown Campus, Penrith, New South Wales, 1797.

PhD and research students' evaluations of supervision were collected across Australian universities to provide an extensive assessment of the quality of research supervision and appropriateness of research facilities. Here we evaluate issues, complexities, challenges, and appropriateness for using such ratings to make benchmarking comparisons between different universities and programs.

Two versions of the Postgraduate Research Experience Questionnaires (PREQ), a multidimensional measure of PhD and research Masters students' evaluation of the quality of research supervision, were administered to recent graduates (n=1832) from 29 Australian and 3 New Zealand Universities. At the level of the individual student, responses had reasonable psychometric properties. Consistent with a potential use of these instruments to benchmark the quality of supervision across all Australian universities, the present study evaluates the extent to which responses reliably differentiate between universities, academic disciplines, and disciplines within universities. Based on fitting two-level (individual student, university) and three-level (individual student, discipline, university) multilevel models to the data, the responses failed to differentiate among universities, or among disciplines within universities (although there were small discipline differences across universities). The results demonstrate that responses that are adequately reliable at one level (individual student) may have little or no reliability at another level (university). We conclude that PREQ responses should not be used to benchmark Australian universities or disciplines within universities. Furthermore, we argue that PREQ responses, as presently formulated, are unlikely to be useful for most other conceivable purposes.

Universities throughout the world are undertaking benchmarking exercises in which they compare themselves to other universities on appropriate indices in order to establish their current levels of performance and to initiate continuous self-improvement [see overview by McKinnon, Walker, & Davis, 2000, for a description of Australian benchmarking; related information from other countries can be found through relevant websites for the United Kingdom (www.niss.ac.uk/education/qaa/), NZ (www.aau.ac.nz/), and the USA (www.chea.org/)]. In order to pursue benchmarking exercises, there is a need for a comprehensive set of benchmark indicators that: focus on outcomes; measure functional effectiveness rather than superficial criteria (i.e., are chosen because they are easily "countable"); are systematically developed so as to have good content (and "face") validity; and differentiate between universities so as to provide appropriate standards as a basis of ascertaining excellence and continuous improvement. In the present investigation we explore some of the issues, complexities, and challenges in attempting to benchmark the quality of research supervision of research and PhD students across a large sample of universities and across similar disciplines in different universities.

Within such a benchmarking framework, it is particularly difficult to establish appropriate outcomes to measure the effectiveness of programs for PhD and postgraduate research students. Even at the undergraduate teaching level where there is widespread use of students' evaluations of teaching effectiveness, there is a limited basis for making comparisons across universities or across similar academic departments from different universities. At the PhD level, there is little research into the systematic use of student surveys to evaluate the quality of PhD research supervision, and, apparently, no research that attempts to compare effectiveness across large numbers of different universities. Within the broader context of a benchmarking exercise, there is a need for the examination of substantive issues relevant to the evaluation of research student supervision, the development and evaluation of an appropriate survey instrument for collecting the data, and methodological issues associated with the appropriate analysis of such data. Hence, within the broader issue of evaluating a large-scale experiment in benchmarking, the purposes of the present investigation are to: (a) describe the extensive development of the Postgraduate Research Experience Questionnaire (PREQ) that is designed to measure the extent to which PhD and postgraduate research students have satisfactory experiences in relation to the quality of their research supervision; and (b) to evaluate the usefulness of ratings by PhD and postgraduate students' evaluations of their postgraduate research and supervisory experience (hereafter referred to as PhD students' evaluations) for making benchmark comparisons across universities and academic disciplines (see footnote 1). Because of the dearth of previous research in this area and the scope of the present investigation, the results should be of broad relevance to PhD-granting universities throughout the world.

The multilevel modeling perspective demonstrated in this study is also important for higher education research (see Ethington, 1996). Almost all data in higher education are inherently multilevel, even though this feature of the data is typically ignored. Depending on the application, the different levels of analysis might include countries, geographic regions or states, different universities, faculties or departments within universities, and individual students or academic staff. As illustrated in the present investigation, research, policy questions, data, and statistical analyses that are appropriate at one level of analysis may be inappropriate or even misleading when evaluated at another level of analysis. Although Ethington's (1996) handbook chapter clearly establishes the relevance of this methodological approach in higher education research, there are few examples of substantive studies in higher education that have used it so that the present investigation helps fill this gap between appropriate and typical research practice.

The substantive issue – the effectiveness of research supervision and the use of PhD students' evaluations as an indicator of this effectiveness – is our main focus and an important concern in higher education. Although there is a vast research literature on undergraduate students' evaluations of classroom teaching effectiveness and some research on the quality of supervision of research and PhD students (e.g., Anderson & Swazey, 1998; Hockey, 1995; Holdaway, 1996; Pearson, 1996), there is little research on the reliability and validity of PhD students' evaluations. For this reason, we begin with a review of the use of student ratings to evaluate teaching effectiveness where there is an extensive research literature and well-established results that are relevant to an evaluation of PhD students' evaluations.

Students' Evaluations of Teaching Effectiveness

In higher education, there is a long history of research and much debate into the use of students' evaluations of teaching effectiveness (e.g., d'Apollonia & Abrami, 1997; Feldman, 1997, 1998; Greenwald & Gillmore, 1997; Marsh & Roche, 1997; 2000; McKeachie, 1997a, 1997b). Effective teaching is a hypothetical construct for which there is no adequate single indicator. Hence, the validity of students' evaluations of teaching or of any other indicator of effective teaching must be demonstrated through a construct validation approach. Extensive reviews of this research (e.g., Abrami, d'Apollonia, & Cohen, 1990; Cashin, 1988; Cohen, 1980; Feldman, 1989a, 1989b, 1997, 1998; Marsh, 1984, 1987; Marsh & Dunkin, 1997, McKeachie, 1979, 1997a, 1997b) have consistently shown that, with careful attention to measurement and theoretical issues, students' evaluations of teaching are: 1) multidimensional; 2) reliable and stable; 3) primarily a function of the instructor who teaches a course rather than the course that is taught; 4) relatively valid against a variety of indicators of effective teaching; 5) relatively unaffected by a variety of variables hypothesized as potential biases, such as expected course grades, class size, workload and prior subject interest; and 6) demonstrably useful in improving teaching effectiveness when coupled with concrete enhancement strategies in specific areas that teachers target for improvement. Emphasizing the individual teacher (or class-average) as the appropriate unit of analysis for students' evaluations of teaching effectiveness, Marsh and Roche (1997; also see Marsh, 1987) stressed that analyses must be conducted at the appropriate unit of analysis in relation to the intended use of the ratings. This student evaluation research provides one model of an ongoing research program to evaluate the reliability, stability, factor structure, construct validity, potential biases, and usefulness for improving practice based on PhD students' evaluations of their supervisors.

Unit of Analysis Problem

The appropriate unit of analysis is a critical methodological issue in student evaluation research that has particular relevance to our evaluation of PhD students' evaluations. Fortunately, however, there is a clear consensus in student evaluation research that the class-average or individual teacher is the appropriate unit of analysis rather than the individual student (e.g., Marsh, 1987). Thus, support for the construct validity of student evaluation responses can only be demonstrated at the class-average level and the reliability of responses is most appropriately determined from studies of interrater agreement that assess error due to the lack of agreement among different students within the same course (see Gilmore, Kane, & Naccarato, 1978 for further discussion). The correlation between responses by any two students in the same class (i.e., the single-rater reliability; Marsh, 1987) is typically in the .20s. However, the reliability of the class-average responses depends upon the extent of agreement among students within the same class and the number of students rating the class: .95 for 50 students, .90 for 25 students, .74 for 10 students, and .60 for five students. Given a sufficient number of students in any one class (or, perhaps, averaged across different classes taught by the same teacher if the number of student in any one class is less than 20) the reliability of class-average ratings is very good. Similarly, support for the construct validity of student evaluation responses must be demonstrated at the class-average level (e.g., relations with class-average achievement, teacher self-evaluations).

In trying to separate the effects of the teacher and the course, Marsh (1987; Marsh & Dunkin, 1997) reported that the correlation between overall teacher ratings of different instructors teaching the same course (i.e., a course effect) was -.05, whereas correlations for the same instructor in different courses (.61) and in two different offerings of the same course (.72) were much larger. These results support the validity of student evaluations as a measure of teacher effectiveness, but not as a measure of the course quality that is independent of the teacher. Marsh and Bailey (1993) further demonstrated that each teacher has a characteristic profile on the different evaluation factors (e.g., high on organization and low on enthusiasm) that was distinct from the profiles of other instructors and generalized across course offerings over a 13-year period. Although there is some research suggesting discipline differences (e.g., a weak tendency for higher ratings in humanities and lower ratings in sciences; see Centra, 1993), these effects account for very little variance and there is ongoing debate about how these differences should be interpreted. Indeed, in many student evaluation programs, ratings for a given class are "normed" in relation to similar classes (similar in terms of student composition, level, and discipline), implying that such differences may not be important.

Hence, at least for the content of items typically considered in this student evaluation research (e.g., enthusiasm, learning/value, organization, rapport, group interaction, breadth of coverage, examinations), the appropriate unit of analysis is the individual teacher and not the individual student. Although it may be possible to construct an alternative set of items that would capture the quality of a course or program that was reasonably independent of the effects of specific teachers, there is little empirical support for this possibility in the student evaluation literature. This conclusion is particularly relevant for the present investigation of PhD students' evaluations in which our focus is on the overall postgraduate experience at the broad level of the university and disciplines within a university rather than the effectiveness of individual supervisors. Extrapolations from student evaluation of teaching research suggest that there should be considerable variation at the level of individual supervisors if there are a sufficient number of ratings for each supervisor, but little or no variation at the level of the discipline or university. Hence, we aruge that the unit of analysis issue is one of the critical complexities in the appropriate analysis of PhD students' evaluations so that a methodological focus on multilevel modeling is an important component in the evaluation of these issues.

Students' Evaluations Can Lead to Improved Teaching

One intended purpose of PhD students' evaluations is to provide informative feedback that will lead to the improvement of research supervision. There is clear evidence that feedback from students' evaluations of teaching, coupled with appropriate consultation, can lead to improved teaching effectiveness (see reviews by Cohen, 1980; L'Hommedieu, Menges & Brinko, 1990;. Marsh, 1987; Marsh & Dunkin, 1997; Marsh & Roche, 1993). For example, in a study by Marsh and Roche (1993), randomly assigned intervention- and control-group- teachers completed self-evaluations and were evaluated by students before and after the intervention. An essential component of the intervention was a set of teaching strategy booklets – one for each factor on the student evaluation instrument. Teachers selected the factor to be targeted in their individually structured intervention and then selected the most appropriate strategies from a book of strategies relevant to that factor. The intervention teachers improved significantly more than control group teachers. Furthermore, for the intervention group (compared to control group), targeted dimensions improved substantially more than nontargeted dimensions. The study demonstrated that feedback from students' evaluations of teaching and consultation are an effective means of improving teaching effectiveness. It is important to note that this intervention can only be conducted with a well-designed, multidimensional instrument and that the specificity of the intervention effects to the targeted dimensions further supports the construct validity of multidimensional students' evaluations of teachings. The lessons from this research that may be useful for improving the quality of postgraduate supervision based on PhD students' evaluations are that: the feedback needs to be specific to each supervisor; supervisors may need concrete strategies about how to improve their supervision; and this feedback may need to be complemented by a trained consultant. Even when supervisors are motivated to improve their supervision and have feedback about their strengths and weaknesses, they still need professional assistance on how to actually improve their supervision. Because university academics typically receive even less training (and have been exposed to even fewer role models) in how to be effective supervisors than in how to be effective classroom teachers, we expect that these results from the student evaluation literature will generalize to research supervision.

Multiple Level of Effectiveness in School Effectiveness Research

In Australia and throughout the world there is an increasing emphasis on accountability and the need to enhance teacher and school effectiveness in both elementary and secondary schools (e.g., Hill & Rowe, 1996, 1998; Rowe & Hill, 1998; Rowe, Hill & Holmes-Smith, 1995; Rowe & Rowe, 1999; Scheerens & Bosker, 1997). Influenced by early production-functions and economic rationalist perspectives, much of this research has focused almost exclusively on academic achievement as an outcome measure. Furthermore, it typically has not taken into account extreme input differences (i.e., if entering students are 1 *SD* above the mean of standardized achievement tests upon entering a school but only 0.5 *SD* above the mean when leaving, then, perhaps, the school has not been 'effective'). More recently, sophisticated statistical procedures incorporating structural equation modeling and multilevel modeling have provided more defensible indices of growth or 'value-added' gains over time that can be attributed to a school, and more particularly to within-school class/teacher effects (e.g., Rowe, 1999; Rowe & Hill, 1998; Rowe, Hill & Holmes-Smith, 1995; Rowe & Rowe, 1999; Rowe, Turner & Lane, in press). In contrast to earlier research that did not account for the inherent hierarchical structure of the data, this more recent research has clearly demonstrated that effective schools are primarily a function of effective teachers within these schools. Once class/teacher effects have been taken into account (45% to 59% of the variance), there is little residual variance at the school level (< 10%). Furthermore, even in the most effective schools there is substantial variation at the class/teacher level. Indeed, Monk (1992) cites a number of studies in support of the observation that: "One of the recurring and most compelling findings within the corpus of production function research is the demonstration that how much a student learns depends on the identity of the instructor to which that student is assigned" (p. 320).

Findings from school effectiveness research also clearly demonstrate that it is important to control input differences before interpreting outcome differences (i.e., a value-added model), and have prompted a major reassessment of knowledge about educational effectiveness in terms of teaching and learning outcomes. This is relevant to PhD students' evaluation research to the extent that there are university-to-university differences in the initial quality of students enrolling in different universities. More generally, whereas this school effectiveness research comes from a very different perspective than research on students' evaluations of university teaching as reviewed earlier, both research literatures lead to a similar conclusion that the individual teacher – or individual supervisor in the case of research supervision – is the most important unit of analysis in assessing the quality of education.

Postgraduate Research Experience Questionnaire (PREQ): Development and Preliminary Evaluations

In order to explore the issues, complexities, and challenges in the use of PhD students' evaluations to benchmark universities or disciplines with universities, it is important to have an appropriate survey instrument. Here, we briefly describe the background to the development and evaluation of the PREQ that was the source of data in the present investigation.

Australia, like many other countries, is seeking ways to improve the quality and enhance the accountability of its higher-education sector (e.g., Harmon, 1999). Consistent with this aim the Australian government and Australian universities have cooperated to collect standardized data that can be used to compare outcomes of any one university with those across all universities or with those of similar universities – a benchmarking exercise. Such comparisons have provided valuable information about research publications, research funding, undergraduate teaching, and a variety of other indicators of effective universities. Thus, for example, highly standardized and audited measures of research productivity – peer-reviewed publications and research funding – are used to rank Australian universities and disciplines within universities and to determine, in part, the research infrastructure funding that different universities receive.

In 1991 the Australian government commissioned trials of the Course Evaluation Questionnaire in order to monitor the quality of students' university experiences. The Course Evaluation Questionnaire is now routinely completed by graduates from all Australian universities within a few months of graduation. The responses assess characteristics of good teaching and effective learning such as enthusiasm, feedback, clarity of explanations, the establishment of clear goals and standards, the development of generic skills, the appropriateness of the workload and assessment, and an emphasis on student independence (Ainley & Long, 1994; Ramsden, 1991). The intent of the Course Evaluation Questionnaire is to provide an overall perspective of student experience and students are instructed to think about their educational experience as a whole rather than to identify specific subjects, classes, or teachers. Results of this exercise are broadly available, for example, through *The Good Universities Guide to Australian Universities* (Ashenden & Milliken, 1995) that has extensive comparative data that are used widely by potential students to select universities. Despite widespread use of the Course Evaluation Questionnaire, its inappropriateness was broadly recognized for purposes of evaluating the quality of supervision of postgraduate research (PhD and research masters) students. This led to the development and evaluation of the PREQ instrument to measure PhD students' evaluations.

In 1996 the Australian Department of Employment, Education, Training and Youth Affairs commissioned the Graduate Careers Council of Australia to develop the PREQ to measure the experiences of PhD and research higher degree students and the Australian Council for Educational Research (ACER) to evaluate trial results based on this instrument (see ACER, October, 1999, for a more detailed description of the rationale, development, and evaluation). The intended purposes of the PREQ were to provide a multidimensional measure of the experience of PhD and research higher degree graduates, to provide comparative information that would allow identification of centres of good practice and to assist institutions with below-average ratings, and, perhaps, to inform students where they were likely to receive good support for pursuing a research higher degree.

An advisory board was established that represented the survey management group and had broad representation of the higher-education sector including academics, research administrators, and research students. The process used to develop the PREQ incorporated extensive input from diverse sources that included: reviews of current literature; current institutional evaluation research; good practice and outcomes in relevant areas; existing instruments used in different universities; feedback from higher education staff with relevant experience; a special conference to explore the development of the instrument; and focus groups with current research higher degree students (ACER, 1999). Based on broad input, the advisory group developed a list of items and issues to be covered on the PREQ. These were then tested with two focus groups of research higher degree students and, concurrently, sent for comment to academic staff with appropriate expertise. This led to the development of a list of potential items to be included. In a third focus group, students completed the items and commented on the items, the instructions, and the format of the instrument. This led to minor rewording of some items and the accompanying instructions. This was followed with further consultation with the advisory group and development of the final pilot version of the PREQ instrument. This systematic development of the PREQ instrument provides strong support for its content validity and the "face validity" of the responses from the perspectives of the major stakeholders.

For purposes of trialing the PREQ, all Australian universities were invited to participate in the pilot studies and most agreed to do so (see ACER, 1999, for list of participating universities). In order to evaluate the most appropriate response format, two substantially similar instruments were developed based on "agree-disagree" and "satisfied-unsatisfied" response scales. Both instruments were considerably longer than the eventual form was intended to be, allowing for the

Paper Presented at the Joint AARE/
NZARE Conference, Auckland, 2003
Australian Association
for Research in Education

selection of the best items. The first data collection was based on PhD and masters research students who graduated between 1 October 1996 and 30 September 1997. Universities mailed survey forms to their recent graduates of research higher degree programs and followed up non-respondents. A second data collection was conducted in 1998 for students graduating in the following year. Across both data collections, there were responses from 1832 research higher degree students representing 29 Australian and 3 New Zealand universities. Analyses included exploratory factor analysis, confirmatory factor analyses, reliability analysis, item response theory analyses, and an evaluation of missing and not appropriate responses that were used to develop the final set of items. The factor analyses demonstrated that both versions of the PREQ resulted in five evaluation factors (Supervisor, Climate, Clarity, Infrastructure, and Skill Development), whereas the agree version resulted in a sixth factor about the thesis examination process (Thesis). The items used to infer each of these scales (ignoring for now factor analyses and reliability results that are presented in subsequent discussion) are presented in Table 1.

In summary, the development of the PREQ instrument supported its content (and face) validity, whereas results based on two large data collections from PhD students from most Australian universities supported its psychometric characteristics (reliability and factor structure) based on responses by individual students. On the basis of this evaluation, it was recommended that the "agree" version of the PREQ instrument (see Table 1) should be used (ACER, 1999).

Table 1 **:Confirmatory Factor Analysis Structure For Postgraduate Research Experience Items**

	Factor Loading					
Supervisor (SUP)	SUP	SKL	CLM	INF	EXM	CLR
Supervision was available when I needed it	.81	0	0	0	0	0
My supervisor/s made a real effort to understand difficulties I faced	.85	0	0	0	0	0
My supervisor/s provided additional information relevant to my topic	.80	0	0	0	0	0
I was given good guidance in topic selection and refinement	.78	0	0	0	0	0
My supervisor/s provided helpful feedback on my progress	.87	0	0	0	0	0
I received good guidance in my literature search	.71	0	0	0	0	0
Skill Development (SKL)						
My research further developed my problem-solving skills	0	.77	0	0	0	0
I learned to develop my ideas and present them in my written work	0	.75	0	0	0	0
My research sharpened my analytical skills	0	.78	0	0	0	0
Doing my own research helped me to developed my ability to plan my own work	0	.67	0	0	0	0
As a result of my research, I feel confident about tackling unfamiliar problems	0	.66	0	0	0	0
Climate (CLM)						
The department provided opportunities for social interaction with other postgraduate students	0	0	.67	0	0	0
I was integrated into the department's community	0	0	.81	0	0	0
The department provided opportunities for me to be involved in the broader research culture	0	0	.81	0	0	0
A good seminar program for postgraduate students was provided	0	0	.61	0	0	0
I used the research ambience in the department or faculty to stimulate my own work	0	0	.73	0	0	0
Infrastructure (INF)						
I had access to a suitable working space	0	0	0	.69	0	0
I had good access to the technical support I needed	0	0	0	.80	0	0
I was able to organise good access to necessary equipment	0	0	0	.78	0	0
I was given good access to computing facilities and services	0	0	0	.76	0	0
There was appropriate financial support provided for research activities	0	0	0	.59	0	0
Thesis Examination (EXM))						
The thesis examination process was fair	0	0	0	0	.80	0
I was satisfied with the thesis examination process	0	0	0	0	.96	0
The examination of my thesis was completed in a reasonable time	0	0	0	0	.56	0
Clarity (CLR)						
I developed an understanding of the standard of work expected	0	0	0	0	0	.77
I understood the required level for the thesis	0	0	0	0	0	.86
I understood the requirements of thesis examination	0	0	0	0	0	.76
Factor Correlations (and coefficient α estimates of reliability)						
Supervisor (α = .91)	1					
Skill Development (α = .85)	.43	1				
Climate (α = .85)	.49	.40	1			
Infrastructure (α = .83)	.52	.43	.76	1		
Thesis Examination (α = .77)	.37	.27	.35	.28	1	
Clarity of Expectations (α = .82	.59	.55	.44	.46	.47	1

Note: Analyses were based on the N = 939 sets of responses by individual students to the agree-disagree version of PREQ. Parameter estimates are presented in completely standardized format. Factor loadings of zero (0) are fixed according to the a priori design of the model.

An important use of the responses would be to benchmark Australian universities in relation to PhD students' evaluations such that the results from any one university (or discipline within a university) could be compared to national normative data and to responses from similar universities. Thus, for example, McKinnon, et al. (2000) was commissioned by the Higher Education Division of the Australian Department of Education, Training and Youth Affairs to develop benchmark outcomes for benchmarking Australian universities. Their proposed benchmarks were intended to cover a wide variety of different university functions. They specifically noted, however, that PREQ ratings by research degree students would provide a useful measure of the quality of research training and stated that "It is envisaged that student experience and satisfaction will be benchmarked by implementation of a Postgraduate Research Experience Questionnaire (PREQ) now in trial use" (p. 101, Benchmark 8.4). The main focus of the present investigation is to evaluate the PhD students' evaluations in relation to this potential use of them.

Questions to Be Addressed in the Present Investigation

The overarching focus of the present investigation is on the use of PhD students' evaluations for purposes of benchmarking the research supervision at different universities. Although research reviewed earlier on the psychometric properties of PhD students' evaluations at the level of the individual student is very encouraging, there are several important questions that need to be addressed before it can be concluded that PhD students' evaluations in general, or PREQ responses in particular, are appropriate for making comparisons between universities or disciplines within universities. First, the appropriate unit of analysis for the analysis of PhD students' evaluations must be established and appropriate evaluations need to be conducted at that unit of analysis. Second, in a related concern, PhD students' evaluations must be able to discriminate between different universities if they are to be used to compare different universities.

PhD Students' Evaluations and the Appropriate Unit of Analysis Problems

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

Although analyses of PREQ responses described earlier were mostly based on the individual student as the unit of analysis, we argue *a priori*, as is the case for students' evaluations, that the appropriate unit of analysis for evaluating PhD students' evaluations SHOULD BE the supervisor who is being evaluated. As with the students' evaluations research, analyses conducted at the level of the individual student may be largely irrelevant. This presents a serious problem for the PREQ analyses in that the individual supervisors were not even identified. (There would, however, be potentially serious issues of confidentiality if supervisors were identified and place students in a conflict of interest situation. Unlike evaluations of classroom teaching, any one supervisor is unlikely to have many students completing their PhD in a given year, and students are likely to be dependent on supervisors for letters of reference for at least the early part of their subsequent career).

From this perspective, the appropriate estimate of reliability should have been the extent of agreement among different students rating the same supervisor (inter-rater reliability). This reliability estimate would be more appropriate and is likely to have been substantially lower than the coefficient alpha estimates of reliability reported in the ACER (1999) report. Based on extrapolations from student evaluation research (Marsh, 1987) and the small number of research students supervised by the same supervisor who are likely to graduate in any given year, it is likely that the reliability of the PhD students' evaluations at the level of the supervisor would be unacceptably low (e.g., .5 or less). This question is moot in the present investigation in that students did not identify their supervisor when completing the PhD students' evaluations. A more relevant question – one that can be addressed with the available information and that is consistent with potential applications of PhD students' evaluations – is the reliability of the PhD students' evaluations at the level of the university or discipline within the university. Here also, the reliability depends on the extent of agreement among different students within the same unit (university, or discipline within a university) and the number of students within that unit.

Another perspective on the unit-of-analysis issue may be the target unit of analysis based on the content of the items. Many of the PREQ items, however, are ambiguous in this respect (see Table 1). Whereas the Supervisor factor and perhaps the Skill Development factors are aimed at supervisors, the Climate and Infrastructure factors may refer to either the supervisor or the academic unit, and the Thesis Examination and Clarity factors may refer to the entire university. Hence, the potential confusion about the appropriate unit of analysis is also evident in the construction of the PREQ items. It is, nevertheless, likely that the individual supervisor is critical in all aspects of the supervision and research training experience.

A particularly important concern about the unit of analysis is the target unit at which the ratings are intended to be used. Particularly since individual supervisors were not even identified, the intended unit of analysis appears to be the entire university, or specific disciplines within each university. Indeed, this supposition is supported by the claim PREQ responses are potentially useful for benchmarking exercises.

Methods

Sample and Procedures

The present investigation is an analysis of data collected in the original trial of the PREQ instrument (ACER, 1999; also see earlier discussion of the background leading to the development of the PREQ). The data consist of responses by 1832 students who recently completed a PhD or research masters degree from one of 29 Australian and 3 New Zealand universities. Responses were from one of two student cohorts; those graduating between October 1996 and September 1997, and those graduating the following year. Roughly half the students completed the agreement-disagreement version of the PREQ and half completed the satisfactory-unsatisfactory version of the instrument. In all cases, each university was responsible for mailing copies of the instrument to their recent graduates and following up non-respondents. The average response rate was 45% for the 29 Australian universities and somewhat lower for the three New Zealand universities. Although the overall sample of respondents and those responding to each version of the instrument are not random samples of the entire population of students, comparisons presented by ACER (1999) suggested that they are broadly representative of the population of research students.

Variables

In the present investigation our focus is on the use of PREQ responses as a means to the examination of issues, complexities, and challenges in the application of PhD students' evaluations as a basis for benchmarking universities rather than the evaluation of a particular instrument. From this perspective, however, it is fortunate that the PREQ instrument has good psychometric properties when evaluated at the level of the individual student (see earlier discussion and subsequent preliminary analyses presented in this study) and that there exists a large, national database of PhD students' evaluations that is appropriate for a multilevel analysis.

In the present investigation, all analyses were conducted with scale scores based on the final recommended version of the PREQ (see Table 1). As in the original reports, scale scores were computed as the mean of non-missing responses to the items designed to measure each scale. For purposes of the present investigation, variables considered include: (a) six scale scores and the overall rating based on the agree version of PREQ; (b) five scale scores and the overall rating based on the satisfactory-unsatisfactory version of PREQ; (c) University; (d) Discipline: a narrow classification reported in the original study consisting of 44 different disciplines (37 of which were actually represented in the data collected) and a broad classification of 10 disciplines used in the present investigation (Agriculture (6%), Architecture (1%), Humanities (16%), Social Sciences (11%), Business (7%), Education (11%), Engineering (10%), Health/Medicine (10%), Life/Biological Sciences (15%), Physical sciences (12%)); (e) Five Student Characteristics: PhD (52%) vs. Masters research student; Part-time (32%) vs. Full-time; Female (48%) vs. male; NonEnglish Speaking Background (NESB, 25%) vs. English speaking background; Age (orthogonal linear and quadratic contrasts based on five age categories). All students were included who completed at least 75% of the survey items and had complete data for the university, discipline, and four of the student characteristics (for student age, the mean age was assigned for 30 students who left this item blank even though they otherwise had complete data). Based on this selection procedure, responses from 1749 of the original 1832 cases were included in the final analyses. Unreported analyses conducted as part of the present investigation suggest that this exclusion of students with missing data had little or no effect on the results reported here.

Statistical Analyses

Statistical analyses consisted primarily of multilevel analyses conducted with the commercially available *MLwiN* (Goldstein et al., 1998) and LISREL (Jöreskog & Sörbom, 1999) statistical packages. A detailed presentation of the conduct of multilevel modeling (also referred to as hierarchical linear modeling) is available elsewhere (e.g., Bryk & Raudenbush, 1992; Goldstein, 1995; Goldstein et al., 1998; Rowe, 1999). Particularly in educational research, characteristics associated with individual students who are clustered into academic units (e.g., classes, departments, faculties, disciplines, schools), pose special problems for analysis. These include the appropriate levels of analysis, aggregation bias, heterogeneity of regression, and associated problems of model misspecification due to lack of independence between measurements at different levels. Thus, it is inappropriate to pool responses of individual students without regard for other levels unless it can be shown responses in the higher-level academic units do not differ significantly from each other. Moreover, reliability estimates, factor structure, and relations with external criteria observed at one level might not bear any straightforward connection to relations observed at another level. Multilevel analyses allow researchers to simultaneously consider multiple units of analysis within the same analysis. In the variance component models, estimates of the variance (and tests of statistical significance) at each level (e.g., individual student and university) are determined. In subsequent models additional predictor variables (e.g., student characteristics) are added to determine their effects and their influence on variance components.

A critical complication in the present investigation is how to compare results across the satisfactory and agree versions of the PREQ instruments. Although there is substantial overlap in the factors measured by each version, the actual wording of the items and response scales are different. For some analyses we merely conducted separate analyses for responses to each version. We also, however, developed an alternative approach. Both versions of the instruments contain a similarly worded overall rating item. For purposes of the present investigation, we normalized and standardized the data using the “normal” procedure available in LISREL (Jöreksog & Sörbom, 1999). This transformed responses to each item to a common metric with a common mean and SD (based on the untestable but plausible assumption that the “true scores” for the two groups are actually the same). In this way, we constructed an overall rating score that was common to students completing both versions of the instruments that allowed us to conduct analyses across the entire sample of students. This is important, because it allows testing more formally whether the effects (e.g., the student characteristics, discipline) differ significantly for the two groups. Of particular relevance to the present investigation, this strategy also increases the statistical power of tests of the variance associated with higher levels (e.g., the university or discipline within the university) where the sample sizes in some cases are low. Although this strategy is based on untestable assumptions about responses in the two groups, a detailed comparison of the parallel analyses of the overall rating item in each group separately with those based on the combined overall rating item across the two groups (see subsequent discussion) supports this approach.

The selection of the appropriate units of analysis for the present investigation is not straightforward. In all analyses to be presented, the individual student is always the lowest level whereas the university is always the highest level. In all cases, both these effects are appropriately considered to be random effects (see Bryk & Raudenbush, 1992; Goldstein, 1995; Rowe, 1999). The classification of academic discipline as random or fixed, however, is more complicated. If the concern is to test for the statistical significance of differences between a fixed set of broad disciplines, then it might be appropriate to consider these as fixed effects. On the other hand, if the concern is with variance associated with particular groups of students associated with a narrowly defined discipline within each university who might have postgraduate experiences that differ systematically from those of students in other disciplines within the same university, then it might be more appropriate to consider discipline as a random variable. In the present investigation, we sought a compromise strategy to pursue both possibilities.

1. First, we fitted a set of two-level models (level 1 = students, level 2 = university). In Model 1 (variance component model), no fixed effects were included. In Model 2, the fixed effects of student characteristics were included. In Model 3, the additional fixed effects of the broad discipline classification (with 10 categories) were included. The 10-category classification of discipline was represented by 9 dummy dichotomous (0,1) variables in which the "left-out" discipline was the discipline that received lowest overall ratings (Humanities). Hence, the test of statistical significance for each remaining discipline was a test of whether ratings were significantly higher than those in Humanities (see Bryk & Raudenbush, 1992; Goldstein, 1995; Rowe, 1999). Under appropriate conditions, the change in likelihood ratio associated with the introduction of new variables in each model can be used to assess the improvement of fit due to the introduction of all fixed effects entered at that level. In preliminary analyses, we also ascertained that the narrow (a 37-category) classification of disciplines was not able to explain significantly more variance than the broad (10-category) classification of disciplines. For this reason, we chose to use the more parsimonious classification to represent the fixed effects of discipline.

2. Second, we fitted a parallel set of three-level models (level 1 = students, level 2 = narrow discipline, 3 = university). As in the two-level analyses, we evaluated three models that included no fixed effects (Model 1), student characteristics (Model 2), and the broad (10-cateogory) classification of disciplines. This final model in which discipline is included as both a fixed effect and a random effect is justified, we argue, because of the competing interpretations of discipline. A detailed comparison of the results from parallel analyses of fitting the two- and three-level models to the data provided support for the appropriateness of conclusions based on this final model.

Preliminary Analysis of Psychometric Properties at the Level of the Individual Student. Although the main focus of our study is on multilevel analyses, it is useful to report briefly the psychometric properties of PREQ responses based on analyses of responses by individual students like those traditionally used to evaluate survey instruments. This is relevant in order to show that this is an appropriate instrument in relation to traditional psychometric properties conducted at the level of the individual students and, thus, appropriate for evaluating the overarching issues which are the focus of the present investigation. For these purposes we considered responses to the 27 PREQ items from the "agree" version of the instrument that was recommended in the evaluation of the instrument (ACER, 1999; see p.57). Even though the six scales are relatively short (varying from 3 to 6 items; see Table 1), the coefficient alpha estimates of reliability are good, varying from .77 to .91 (median = .84; Table 1). Exploratory and confirmatory factor analyses each provided strong support for the a priori six factors. A highly restrictive confirmatory factor analysis model was tested in which each item was allowed to load on one and only one factor and residual variance (measurement error) associated with each item was required to be uncorrelated with residual variance components associated with other items. The fit of the model was good in relation to traditional guidelines of goodness of fit (Tucker-Lewis index = .93; relative noncentrality index = .94; and residual mean square error of approximation = .05). The solution was well defined in that the factor loadings relating items to their a priori factor were all substantial and statistically significant (see Table 1). Hence, these preliminary analyses support psychometric properties of the PREQ responses when evaluated at the level of the individual student (for further discussion, see ACER, 1999).

Results

The central questions in the present investigation are whether PhD students' evaluations are able to discriminate between universities or to discriminate between the same discipline across different universities. We address these questions with two sets of multilevel analyses.

Differences Between Universities: Two-level Analysis

University Effects.

In two-level models (level 1 = student, level 2 = university; see Tables 2 and 3) we evaluated: (a) Variance associated with differences between universities (Model 1); (b) Effects due to student characteristics (Model 2), and (c) Effects due to different disciplines (Model 3).

Table 2: **Differentiation Among Universities and Disciplines: Variance Components**

		University Random		University & Discipline Random		
Scale	Model	Univer	Student	Univer	Discip	Student
Agreement						
Overall	1	.031	.930*	.024	.039	.895*
	2	.021	.906*	.017	.029	.880*
	3	.022	.893*	.019	.023	.873*
Supervisor	1	.020	.853*	.016	.030	.826*
	2	.018	.835*	.015	.028	.810*
	3	.019	.818*	.017	.012	.807*
Skill Devel	1	.006	.281*	.006	.000	.281*
	2	.002	.274*	.003	-.001	.274*
	3	.002	.270*	.003	-.002	.272*
Climate	1	.020	.808*	.012	.072*	.743*
	2	.011	.761*	.008	.050*	.714*
	3	.010	.753*	.008	.039	.716*
Infrastruct	1	.030	.791*	.018	.091*	.707*
	2	.011	.738*	.002	.066*	.678*
	3	.009	.716*	.002	.048*	.674*
Thesis Exam	1	.048*	.876*	.046*	.022	..859*
	2	.032	.854*	.031	.006	.849*
	3	.030	.843*	.029	.002	.842*
Clarity	1	.025*	.499*	.024	.020	.481*
	2	.013	.478*	.013	.014	.464*
	3	.013	.474*	.013	.009	.466*
Satisfaction						
Overall	1	.005	.494*	.004	.017	.479*
	2	.000	.485*	.000	.016	.469*
	3	-.001	.476*	-.001	.006	.471*
Supervisor	1	.003	.405*	.002	.010	.396*
	2	.001	.400*	.000	.008	.394*
	3	-.001	.392*	-.001	-.006	.398*
Skill Devel	1	.005	.228*	.003	.008	.221*
	2	.003	.221*	.002	.007	.214*
	3	.003	.216*	.002	.001	.215*
Climate	1	.007	.398*	.006	.050*	.356*
	2	.004	.388*	.003	.038*	.356*
	3	.004	.375*	.003	.018	.359*
Infrastruct	1	.007	.393*	.005	.027*	.369*
	2	.003	.386*	.001	.020	.366*
	3	.002	.370*	.002	.001	.370*

Clarity	1	.007	.419*	.007	.005	.415*
	2	.003	.411*	.003	.002	.410*
	3	.002	.406*	.002	-.007	.412*
Combined Overall[a]						
Overall	1	.018	.975*	.013	.038	.941*
	2	.011	.956*	.007	.032*	.927*
	3	.009	.944*	.007	.017	.929*
	4	.008	.932*	.006	.018	.917*

Note. Two sets of analyses were conducted, two-level analyses in which only the university (n=32) was random and three-level analyses in which both the university and discipline (n=37) within the university were random. For both sets of analyses, three models were evaluated: Model 1 in which there were no fixed effects; Model 2 in which student background variables were estimated as fixed effects; Model 3 in which student background variables and broad discipline (10 classification) were estimated as fixed effects. In addition, for the combined overall analysis only, interaction terms were included to determine if the fixed effects differed for students who completed the two instruments (see Table 3).

[a] The Combined analysis was based on the overall rating item from each of the two instruments (after responses from each instrument were normalized and standardized separately to put them into the same metric).

Table 3: **Fixed Effects of Student Background Variables and Disciplines For Agreement, Satisfaction, and Combined Responses**

	Student Background Variables						Broad Disciplines								
Scale	PhD	PT	Female	NESB	Age-L	Age-Q	Agri	Arch	SocS	Bus	Educ	Engr	Hlth	LifeS	PhyS
Agreement															
Overall	.18*	-.08	-.22*	.12	-.01	-.01	.33*	.11	.01	.38*	.08	.06	.24	.09	.16
Supervisor	-.01	-.01	-.12	.28*	.01	.00	.05	-.07	.01	.20	.08	-.22	.02	-.22	.06
Skill Devel	.22*	-.02	.03	.02	-.04	-.01	.08	-.17	.02	.10	.11	-.05	.01	-.05	-.09
Climate	.19*	-.15*	-.25*	.20*	-.02	.02	.20	-.02	-.01	.17	.06	-.01	.19	.12	.24
Infrastruct	.26*	-.18*	-.22*	.14*	-.04	-.02	.49*	.32	.27*	.50*	.29*	.38*	.47*	.43*	.47*
Thesis Exam	.21*	-.02	-.05	.05*	.07	.01	.02	-.09	.17	.22	.28*	-.10	.17	.06	.12
Clarity	.30*	.07*	-.04	.16*	.03	.00	.11	-.08	.06	.18	.17	.08	.07	.08	-.03
Satisfaction															
Overall	.08	.00	-.12*	.00	.10*	.02	.32*	-.15	.09	.15	.24*	.14	.12	.15	.26*
Supervisor	.03	.05	.02	.03	.06*	.01	.21	.01	.14	.07	.30	.08	.01	.00	.15
Skill Devel	.16*	-.02	.03	.02	-.04	-.01	.08	-.17	.02	.10	.11	-.05	.01	-.05	-.09
Climate	.08	-.03	-.14*	-.03	.01	.03	.09	-.19	.12	.21*	.28*	.23*	.23*	.21*	.36*
Infrastruct	.12*	.02	-.09	.02	.01	.01	.27*	-.07	.15*	.07	.21*	.32*	.25*	.19*	.42*
Clarity	.09	.05	-.03	.05	.08*	.01	.20	-.08	.06	.06	.25	.04	.07	.07	.16
Combined Overall															
Overall	.16*	-.07	-.19*	.05	.06	.01	.39*	.04	.09	.32*	.24*	.11	.25*	.16	.28*
Interact	.03	-.02	-.02	.03	-.09*	-.04	-.01	.02	-.02	.02	-.04	-.02	.01	-.03	-.03

Note. PhD (1=PhD, 0 = Masters) PT (1=Part Time, 0 = Full Time), Female (1=female, 0 = male), NESB (1=NonEnglish Speaking Background), Age-L = Linear effect of age, Age-Q = Quadratic effect of age, Agri = Agriculture, Arch = Architecture, SocS = Social Sciences, Bus = Business, Educ Education, Engr = Engineering, Hlth = Health, LifeS = Life/Biological Sciences, PhyS

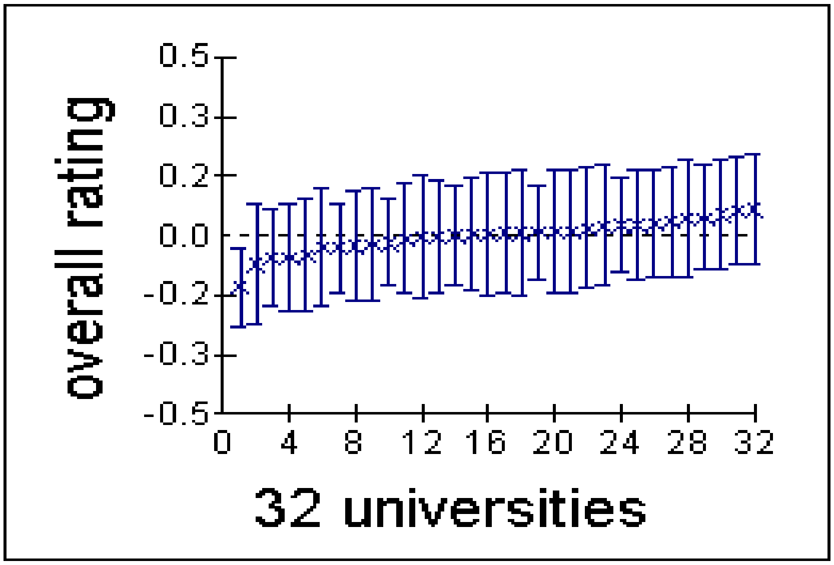
Fixed effects and their statistical significance are based on two-level model in which level 1 = individual student, level 2 = university, but the results were nearly identical for the three level model (that also included the narrow discipline as a random effect).

[a] The Combined analysis was based on the overall rating item from each of the two instruments (after responses from each instrument were normalized and standardized separately to put them into the same metric). In an additional model, interaction terms were also included to determine if the effects based on the overall ratings for one group differed from the other group. Only the linear effect of age differed for the two groups (age was more positively related to overall ratings for students who completed the agreement instrument than those who completed the satisfaction instrument).

* $p < .05$

For the overall ratings based on each version of the PREQ and the combined group, the variance component associated with differences in university was not statistically significant for any of these three models. These results for the combined analyses are illustrated in Figure 1. For purposes of this illustration, the 32 universities are ranked from lowest to highest and an error bar (range of probable error) is placed about the mean rating for each university (for further discussion of this mode of presentation, see Goldstein, et al., 1998; Rowe, 1999). The range of probable error for 31 of the 32 universities includes 0 (the mean for all universities). There is only one New Zealand university that has a mean that differs noticeably from the means of the other 32 universities (see footnote 2). In summary, analyses based on the overall student ratings show that there are no significant differences between universities.

Figure 1: **Two-Level Model. Mean Overall Rating (± 1.96 SD) for 32 universities ranked from lowest to highest. Based on two-level model (level 1 = student, level 2 = university) with student background and broad discipline as fixed effects (Model 3 with university random in Table 2).**



We now turn to an evaluation for the specific evaluation factors: Supervision, Climate, Clarity, Infrastructure, Skill Development, and, for the agree version, Thesis Examination. For the satisfaction version of PREQ, all 18 variance components associated with university differences (6 scales x 3 models) are non-significant. For the agreement version, however, there are marginally significant variance components with Model 1 (.01 < p < .05) for Thesis Examination and Clarity. These differences, however, were nonsignificant when student characteristics were controlled in Model 2 and remained nonsignificant in Model 3 that also controlled for discipline differences. Hence, apparent differences between universities were confounded with student characteristics (e.g., students in PhD and masters programs) so that once these differences were controlled, there was no significant variance associated with differences between universities.

In summary, these results provide no support for the claim that PhD students' evaluations are able to differentiate between different universities. More specifically, there is no significant variation between universities for any of the PREQ scales, or the overall ratings for either versions of the instrument. For purposes of distinguishing between different universities, the reliability of PhD students' evaluation responses does not differ significantly from zero.

Differences Due to Student Characteristics and Academic Discipline

The fixed effects due to student characteristics and academic discipline (Table 3) for the overall ratings are small and largely nonsignificant. In combination, these variables explain only about 5% of the variance in the student overall ratings of their postgraduate research experience based on separate analyses of the agree and satisfaction versions of the PREQ. In each of the analyses, females rate their overall postgraduate experience significantly lower than males. Also, older students tend to give higher ratings for the satisfaction version of the PREQ but not the agree version of the instrument, whereas PhD students tend to give higher ratings on the agree version of PREQ than students completing a research masters degree. The effects of disciplines are mostly small, with only 2 and 3 significant effects in the separate analyses of the agree and satisfaction versions of the instrument. Although the sizes of the discipline effects are similar for the combined ratings, there are more statistically significant effects (5) due to the substantially larger sample size in the combined group. Due to the use of dummy coding (in which the discipline with the lowest overall ratings, humanities, was the base discipline), each of the regression coefficients for the remaining disciplines is a test of whether ratings in that discipline differ significantly from those in the humanities. For the combined overall ratings, ratings are significantly higher than in humanities for 5 disciplines (agriculture, business, education, health sciences, and physical sciences) and do not differ significantly from those in humanities in 4 remaining disciplines (architecture, social sciences, engineering, and life sciences). It is important to reiterate that these discipline differences are ones that generalize across the set of 32 universities and are not specific to particular universities.

A potentially important feature to these results is the inclusion of 15 interaction terms to determine whether any of the student characteristics or discipline effects differs significantly for students who completed the two versions of the PREQ. An omnibus test for all 15 effects is the difference in the likelihood ratios for Models 3 and 4 in Table 2 (22.8) relative to the difference in df (15, one for each of the 15 interaction terms). Because this difference is non-significant ($p > .1$) there is no evidence that any of these fixed effects differ for the two groups. An evaluation of each of the interaction effects (Table 3) indicates that only one difference is significant at a nominal $p < .05$ level. As noted previously, there is a positive linear effect of age in overall ratings on the agree version of the PREQ, but not the satisfaction version. Given the substantial power of these tests (due to the large sample sizes), the results suggest that effects based on the two groups are reasonably similar – at least in terms of responses to these overall-rating items that are worded similarly in the two instruments.

Of greater interest, perhaps, is the pattern of fixed effects that are specific to different PREQ scales and whether the pattern of results provides any support for the construct validity of the PhD students' evaluations (Table 3). For the effects of the student characteristics, the effects are not particularly consistent across the two instruments. In particular, NSEB students tended to give higher ratings than other students for the agree version of PREQ but not the satisfaction version, whereas part-time students tended to give lower ratings on the agree version but not the satisfaction version. Also, age is positively correlated with ratings for two of the satisfaction scales, but not for any of the agreement scales. Results for discipline differences are somewhat more consistent across the specific scales. In particular, the significant effects are associated mostly with the Climate and Infrastructure scales. Even here, however, there is a difference between the two versions in that there are significant effects of discipline for Infrastructure and

Climate ratings with the satisfaction version of PREQ, but only for Infrastructure ratings with the agree version.

Because the wording of items in the specific scales is different in the two versions (as well as differences in the non-randomly assigned samples and differences in the agree and satisfaction response scales), it is difficult to interpret these effects of student characteristics. Because the patterns of results vary across two versions of the PREQ that seem to measure similar constructs, the results suggest that these fixed effects associated with student characteristics may not be particularly robust. In contrast, however, the fixed effects were very similar when based on the two overall rating items that were similarly worded. Hence, we suspect that some of the differences due to apparently similar scales is due to differences between items comprising each scale.

Differences Between Universities and Disciplines Within Universities: Three-level Analysis. In the three-level models (level 1 = student, level 2 = discipline, level 3 = university; see Table 2), the narrow discipline (37 category) classification is added as the second level (with university as the third level). Again, three models were considered with: (a) Variance associated with difference between universities (Model 1); (b) Effects due to student characteristics (Model 2); and (c) Effects due to different disciplines (Model 3). As with the two-level models, we begin with an evaluation of the overall ratings for each version of the PREQ instrument and the combined ratings. Again, however, there are no significant variance components associated with university or discipline for any models based on these overall ratings (Table 2). This lack of differentiation is also evident in the graph of the mean ratings with error bars for each of the 32 universities. As with the earlier results based on two-level models, this graph shows that there is still only the one NZ university that has a mean rating significantly different from the mean across all universities or the mean of any of the other 31 universities.

For analyses based on the specific PREQ factors – also consistent with results based on the two-level models – almost all of the variance components associated with university differences are non-significant. There are, however, several significant variance components associated with discipline for the Climate and Infrastructure factors based on responses to both instruments. It is important to reiterate that discipline variance components reflect discipline differences within universities. These within university differences, however, are potentially confounded with across university differences associated with each discipline. Hence, it is not surprising that the PhD students' evaluation factors showing significant variance components for discipline are also the PhD students' evaluation factors where there are significant fixed effects associated with discipline (see Table 3). Thus, when the fixed effects of discipline are added to the three-level analyses (the three-level Model 3 in Table 2), there is only one variance component associated with discipline that remains marginally significant ($.05 < p > .01$). Hence, although there are discipline differences associated with a few of the evaluation factors, apparently these differences generalize across universities. Although we have not presented the fixed effects for the three-level model, they are nearly identical to those presented for the two level model in Table 3 (i.e., the pattern of significant and nonsignificant differences is exactly the same for the two sets of results). This similarity in the fixed effects based on the two sets of models is not surprising since the variance components associated with university and discipline within university are all very small and mostly nonsignificant (Table 2).

Discussion and Implications

Usefulness of PhD Students' Evaluations for Benchmarking Universities

The most salient finding of this study is that PhD students' evaluations do not vary systematically between universities, or between disciplines within universities. This has critically important methodological and substantive implications for the potential usefulness of PhD students' ratings. Because there is no significant variation at the university level, it follows that the ratings are completely unreliable for distinguishing between universities. This clearly demonstrates why it is important to evaluate the reliability of responses to a survey instrument in relation to a particular application and the level of analysis that is appropriate to their intended use. Although it could be argued that PhD students' ratings were reliable at the level of individual students (e.g., Table 1) these results are not relevant for the likely application of the ratings to discriminate between universities. Substantively, the results of the present investigation place into question the potential usefulness of PhD students' evaluations in benchmarking different universities.

The results of this study also yielded interesting results associated with academic discipline. When the broad set of academic disciplines was included in the multi-level models, there were some significant differences associated primarily with Infrastructure and, to a lesser extent, Intellectual Climate. Thus, for example, Infrastructure support was judged to be significantly better in the science disciplines (e.g., Health, Life and Physical Sciences) than in Humanities. Importantly, however, these results indicated that there were some discipline differences for a few scales that generalized across all universities in the sample. When we fitted a three-level model with discipline as one of the levels (nested under university), the results again indicated that there was some significant variation associated with discipline. For these results, however, two very different interpretations were plausible. This variation could reflect variance associated with specific disciplines within universities, suggesting that there may be some value in benchmarking universities in relation to specific disciplines. Alternatively, this variation associated with discipline might reflect discipline differences that generalized across universities like those already been identified by the fixed effects of discipline (in the two-level analyses). In order to test between these two alternative interpretations, we conducted a final set of three-level models in which we included discipline as one of the levels of analysis and the fixed effects of discipline. These results clearly showed that almost all of the variation due to discipline could be explained by discipline results that generalized across universities, and that almost none could be explained by discipline differences within universities (i.e., the variance components for discipline were nonsignificant for all 5 scales on the satisfaction version of PREQ, 5 of 6 scales on the agreement version of PREQ, and the overall ratings for both versions considered together or separately). These results are important because they clearly demonstrate that PhD students' evaluations are not appropriate for comparing specific disciplines across different universities.

Construct Validity: Potential Criteria Used to Validate PhD Students' Evaluation Responses

Because PhD students' ratings are almost completely devoid of reliability for purposes of benchmarking universities, it also follows logically that they are also invalid for such purposes. In pursuit of concerns about validity, ACER (1999) related PhD students' evaluations to three potential criteria: (a) the Research Quantum (a standardized, externally audited measure of the research productivity of each Australian university compiled by the Australian government); (b) number of Australian Postgraduate Awards (a highly prestigious scholarship available only to the best applicants) obtained by individual universities; and (c) attrition rates in the various postgraduate populations. However, correlations between these external validity criteria and PhD students' ratings of their supervision were not statistically significant, leading to the conclusion that "comparisons based on institution wide research performance are probably inappropriate for establishing external validity" (ACER, p. 78). In contrast, we argue that these external criteria are appropriate for purposes of validating PhD students' ratings in relation to the purpose of benchmarking universities. Indeed, funding to Australian universities for research degree students is based substantially on these criteria on the assumption that they provide surrogate measures of the quality of research supervision. Instead of rejecting the appropriateness of these external criteria, their lack of relation with PhD students' responses calls into question the construct validity of the responses for this purpose. Furthermore, our analyses clearly show that the reason why the responses lack validity in relation to these external criteria is that the responses are completely unreliable at the level of the university. Hence, it is unlikely that PhD students' responses will be meaningfully related to any external validity at the level of the university.

Other Potential Uses of PhD Students' Evaluations.

The focus of the present investigation is on the usefulness of PhD students' ratings of their research supervision for purposes of benchmarking universities. Our results, however, provide strong evidence against the construct validity of the responses, at least for purposes of benchmarking universities. Given this failure, it is appropriate to speculate on other possible uses of the PhD students' responses in their current form. After carefully considering a variety of options, we conclude that they are of limited use for any likely purpose.

If it was ethically, politically, and logistically feasible for students to identify their supervisor and to provide supervisors with this feedback, then their ratings of supervision might be useful for improving the quality of supervision and for administrative decisions that reward effective supervisors. There is support (see earlier discussion) from both student evaluation research at the university level and from school effectiveness research at the elementary and secondary school levels that individual teachers do make a difference. Furthermore, research based on students' evaluations of teaching effectiveness provides a well-established example of how student ratings can be used to improve the effectiveness of individual teachers. For supervisory ratings, however, there are serious impediments to this use of PhD students' evaluations. Students asked to rate the quality of their supervision, knowing that responses would be returned to their supervisor, would have a serious conflict of interest. Because any particular supervisor would typically have no more than one student completing a research degree in a given year – certainly no more than a very few – there could be no effective guarantee of the confidentiality of their responses. Research degree students have already graduated when asked to complete the ratings in the present investigation. However, they are likely to be dependent upon their supervisors for letters of reference to prospective employers for at least the early part of their subsequent career. Furthermore, even if PhD students could be relied upon to give candid responses, the very small number of students evaluating a particular supervisor in any given year (typically none and rarely more than one) would not provide a reliable basis for evaluating the effectiveness of an individual supervisor. Hence, we doubt whether PhD students' evaluations would be particularly useful for evaluating the effectiveness of individual supervisors for either formative feedback or summative personnel purposes.

An additional purpose of PhD students' evaluations might be to provide students with a basis for selecting universities that provide good postgraduate research supervision. Indeed, extensive use of the Course Experience Questionnaire – the evaluation of undergraduate experience that served as one model for the PREQ – is made for this purpose through the publication of these results. The appropriate use of PhD students' evaluations for this purpose, however, also requires that the responses are reliably and validly able to discriminate among universities (or disciplines within universities). Because the present results show that PhD students' evaluations are completely unreliable for this purpose, it would be inappropriate to use the PhD students' evaluations to inform choice of programs.

Averaged across all universities, PhD students' rating of their supervision may reflect the effectiveness of research training across the entire Australian higher-education sector. Even here, however, there are no clearly articulated benchmarks about what constitutes superior, acceptable, or unacceptable responses. Hence, PhD students' evaluations would be of limited usefulness in evaluating the quality of postgraduate research training across the entire university sector. It is also possible that PhD students' evaluations collected over a number of years may provide a benchmark for evaluating system-wide changes in postgraduate training – particularly in a period of much potential change in the Australian postgraduate training policy. Even here, however, there are no standards of what levels are acceptable and it would be difficult to know whether any observed changes represented changes in national policies, university policies, differences in the cohort of students enrolling in research degree programs, or changes in expectations of students over time. Whereas it might be possible to obtain postgraduate student ratings on a very different set of constructs focused more on quality assurance processes that are relevant to university-level policies for purposes of benchmarking universities, the results of the present investigation are not very encouraging. Based on our investigation we conclude that PhD students' evaluations – at least as formulated in this study – are unlikely to be useful for most conceivable purposes.

These results also have important implications for universities that develop their own surveys for use with only their own PhD students. Because PhD students' evaluations do not vary much from university to university, it seems unlikely that the results of such an exercise would be very useful in assessing the quality of supervision at a given university. Furthermore, although there might be small differences associated with particular disciplines on some evaluation factors like those found here, these differences are not easily interpreted. As shown in the present investigation, observed discipline differences are likely to reflect differences that would generalize across PhD students' evaluations of that same discipline across different universities rather than differences that are specific to the particular academic unit within a given university.

Potential Limitations to the Generalisability of Results From the Present Investigation.

We claim that the results of the present investigation are generalizable and have broad relevance to the higher-education research community. Clearly, the broad issues that we address – PhD students' evaluations of the effectiveness of their research supervision and the potential usefulness of these ratings for benchmarking universities – have broad applicability. Had our results provided reasonable support for the usefulness of these ratings as a basis for discriminating between different universities, the procedures would have provided an important model for similar programs in other universities and other countries. Because our results did not support the usefulness of PhD students' evaluations for benchmarking universities, it is important to evaluate the extent to which the nonsignificant results are a function of idiosyncratic features of our study. In particular, it is appropriate to ask the question: Is it likely there would be support for the usefulness of PhD students' evaluations of their supervision in another study based on responses to a different instrument, or responses from a different group of students, or responses from a different set of universities? Although our conclusions must be somewhat tentative, we argue that our results are likely to be broadly generalizable.

An important feature in assessing the generalisability of our results, perhaps, is the PREQ instrument that was the basis of our research. Although our results show that PhD students' evaluations are not appropriate for purposes of benchmarking universities, there is evidence that the PREQ instrument is good according to many traditional criteria when evaluated on the basis of responses by individual students. In particular, the extensive development and refinement of the instrument that involved input from such a diverse group of stakeholders supports its content validity and appropriateness for research students whom completed the instrument. Furthermore, the reliability of responses and particularly the very demanding test based on confirmatory factor analysis provided stronger support for this instrument than most surveys that are used in higher education. This extensive process of development and good psychometric properties justify the use of responses based on this instrument in the present investigation.

Importantly, our results generalized well across the six different PREQ scales and the overall rating item. It may also be reasonable to argue that there are additional scales that might have been included or even that not all of the scales on the current instrument should have been included (although support for the content validity of the instruments may be used to counter these possibilities). However, it seems unreasonable to argue that all of the scales on the current instrument are inappropriate. Furthermore, although PhD students' evaluation is clearly a multidimensional construct, it seems unreasonable to argue against the appropriateness of the overall-rating item. Indeed, for purposes of a benchmarking exercise, it might be relevant to argue that the overall rating is the most important component to consider. In the present investigation there was good generalisability across two somewhat differently worded overall rating items. Because overall rating items that might be included on other instruments are likely to be similar to the ones in the present investigation, our results based on the overall rating items are particularly likely to be replicable across different instruments.

It might be argued that whereas supervisor is the appropriate unit of analysis for many of the PREQ scales, the university might be the more appropriate unit of analysis for one or two of the PREQ scales (e.g., Thesis Examination and Clarity; see Table 1). This follows in that the examination process of the thesis and the expectations of the thesis are largely determined by university-wide policy in Australian universities. To the extent that variance due to the university accounted for substantially more variation in these scales than in the other scales, there would be support for this suggestion. The analyses, however, were consistent in showing that differences between universities were not statistically significant for any of the PREQ scales. Hence, university differences are uniformly nonsignificant, even for PhD students' evaluation factors that might logically be expected to vary from university to university, further undermining support of the construct validity of the PhD students' ratings of supervision.

Significantly, this pattern of results in the present investigation is also consistent with those from the large body of research based on students' evaluations of classroom teaching effectiveness showing that even when students were asked to rate the overall course rather than the overall teacher, their ratings reflected primarily the teacher who taught the course rather than the course that was taught (Marsh, 1987). This continuity with other well-established findings in related areas of research provides good support for the generalisability of results from the present investigation.

The sample of research higher degree students in the present investigation is broadly representative of those graduating from Australian universities. The sample of universities includes almost the entire population of Australian universities and includes a diversity of institutions (e.g., old, well-established research universities and new universities that were primarily teaching colleges and did not grant PhDs prior to the 1990s). Although there are features of the Australian PhD that are different from those in many other countries (e.g., a greater emphasis of the PhD thesis with little emphasis on coursework and qualifying examinations; the use of examiners external to the university to evaluate the thesis), most components of the PREQ are likely to be broadly appropriate across most research university settings. Although it is always desirable to have larger samples, the size of the sample considered here (1832 students from 32 universities) is sufficiently large to address the questions that are the focus of the present investigation. Although it may be possible that substantially larger samples would have resulted in "statistically significant" results due to the increased power of the analyses, it is highly unlikely that the extent of differentiation between universities would be sufficiently large to provide a useful basis for benchmarking universities. Hence, we argue that it is likely that the results reported here will generalize to different university settings.

One particularly relevant limitation of the present investigation that warrants further research is the identification of the supervsior(s) when PhD students rate the effectiveness of their research supervision. Because of this limitation we were unable to evaluate the ability of PhD students' evaluations to differentiate among different supervisors. We argued that there are apparently insurmountable difficulties (e.g., potential conflict of interest and the small number of students for each supervisor) in pursuing this issue. Nevertheless, we also contend that PhD students' ratings would be able to differentiate between supervisors if these difficulties could be overcome. This could be a very important result in terms of improving the quality of research supervision, recognising and rewarding effective supervisors, and, perhaps, matching students to supervisors. It is important to emphasize, however, that even if PhD students' evaluations were able to differentiate reliably between supervisors, this would not undermine our conclusion that the evaluations are not very useful for benchmarking universities. Because the quality of supervision at any given university is likely to be very diverse, it is unlikely that there is substantively meaningful variation in the quality of supervision at the university level. Hence, the limitation of not targeting the particular supervisor in the PhD students' evaluations maybe important for purposes of evaluating the potential usefulness of these ratings for other purposes but apparently is not an important limitation in terms of benchmarking universities.

In conclusion, we argue that the present investigation provides an appropriate starting point for any such evaluations based on different survey instruments, different universities, or different groups of students. In particular, the results demonstrate that traditional (single-level) criteria used to evaluate the psychometric properties of responses to survey instruments may not be appropriate if the purpose of the responses is to compare different universities or academic program within universities. More generally, our methodological approach and results illustrate some of the potential complications and subtleties that are likely to be encountered in such a research program.

A Political Post Script: The Interface Between Academic Research and Political Decision Making

The Australian Department of Education Training and Youth Affairs (DETYA) commissioned the Graduate Careers Council of Australia to develop the PREQ. Preliminary research based on PREQ responses was reported favourably, leading, in part, led to recommendations that the PREQ responses were likely to be useful for benchmarking comparisons. I Professor Marsh (the first author of this article) was asked by the Australian Council of Deans and Directors of Graduate Studies to comment on this preliminary research. He noted concerns along the lines of those documented in this article and, based on these comments, and was requested that he pursue further analyses that formed the basis of this article.

Despite the politically sensitive nature of this undertaking, DETYA and the Graduate Careers Council readily provided the data to conduct the analysis and constructive suggestions. Results of this research were presented at a meeting of the Council of Deans, to a meeting of Australian Pro Vice Chancellors of Research and, subsequently, at the PREQ Seminar that was organised by the Graduate Careers Council in association with DETYA.

Marsh’s conclusions that the PREQ responses did not provide a suitable basis for benchmarking between universities were vigorously debated by the proponents of the instrument at the seminar, but there appeared to be broad acceptance by university representatives that the responses were not appropriate for benchmarking universities. Proponents who were responsible for developing the instrument argued that more data was needed to test the instrument and that these results would be presented at a second PREQ seminar to be held the following year, whereas many other representatives at the seminar felt that it was dubious to proceed further.

In summarising the outcomes of this seminar, a report by the Council of Deans stated: “The question of whether or not attempts will indeed be made to further develop PREQ were left ‘up in the air’. However, it was clear that there is a political agenda around PREQ being pursued by DETYA. Also, there is an apparent determination by the developers to push forward with the development of PREQ and see it in use, despite the very strong message from all but a handful of the participants at the seminar.”

Following concerns that the use of PREQ would proceed despite these reservations (as reported in Campus Review, 19-25 April 2000), the Council of Deans issued a statement (26 April, 2000) rejecting the PREQ as an appropriate basis for benchmarking universities and recommending to appropriate bodies that either the use of PREQ should be discontinued or that universities be advised not to participate in the administration of the PREQ. Shortly thereafter (30 April, 2000), the Australian Vice-Chancellors' Committee (council of Australia’s university presidents) expressing concerns over the possibility that DETYA might ignore conclusions of the first PREQ seminar and pre-empt the likely discussion at a second PREQ seminar.

Following political pressure, DETYA did have a second PREQ seminar in which the participants were carefully selected. I was invited to speak briefly, but was only given access to the data 1 week prior to the meeting. At that meeting, I presented the results based on the new round of data. Even more convincingly than the first round of data, these new results indicated that there were no differences between any two universities. These results strongly replicated the findings presented here. Nevertheless, the meeting rejected the appropriateness of multilevel modeling for these purposes and my conclusions.

DETYA commissioned ACER to do a report on PREQ that was presented at the second seminar. The ACER report (based on single-level analyses) presented results largely similar to mine. However, the ACER report concluded that there were marginally significant differences between the most extreme universities. Nevertheless, even this report cautioned about the inappropriateness of using PREQ to benchmark universities.

The participants of the 2nd PREQ seminar voted to continue to use PREQ.

My results were subsequently subjected to rigourous peer review and published in the Journal of Higher Education, the most prestigious journal in the world in this area of research. (Marsh, H. W., Rowe, K., Martin, A. (2002). PhD students' evaluations of research supervision: Issues, complexities and challenges in a nationwide Australian experiment in benchmarking universities. Journal of Higher Education, 73 (3), 313-348.)

The political organisers of that 2nd PREQ Seminar have largely moved on and I temporarily grew tired of researching politically “hot” topics, but the PREQ remains.

References

Abrami, P. C., d'Apollonia, S., and Cohen, P. A., (1990). Validity of student ratings of instruction: What we know and what we do not. Journal of Educational Psychology, 82: 219-231.

Ainley, J., & Long, M. (1994). The Course Experience Survey: The 1992 Graduates. Graduate Careers Council of Australia, Department of Employment, Education and Training, Canberra, ACT.: Australian Government Printing Service.

Australian Council for Educational Research (October, 1999). Evaluation and validation of the trial Postgraduate Research Experience Questionnaires. :Australian Council for Educational Research.

Anderson, M. S. & Swazey, J. P. (1998). Reflections on the graduate student experience: An overview. New Directions for Higher Education, 26,13-13.

Ashenden, D., & Milligan, S. (1995). Good Universities Guide to Australian Universities, 1996. Melbourne, Victoria: Reed Reference Australia.

Bryk, A.S., & Raudenbush, S.W. (1992). Hierarchical linear models: Applications and data analysis methods. Newbury Park, CA: Sage.

Cashin, W. E. (1988). Student Ratings of Teaching. A Summary of Research. (IDEA paper No. 20). Kansas State University, Division of Continuing Education. (ERIC Document No. ED 302 567).

Centra, J. A. (1993). Determining Faculty effectiveness. San Francisco: Jossey Bass.

Cohen, P.A. (1980). Effectiveness of student-rating feedback for improving college instruction: a meta-analysis. Research in Higher Education 13: 321-341.

d'Apollonia, S. and Abrami, P. C. (1997). Navigating student ratings of instruction. American Psychologist, 52, 1198-1208.

Ethington, A. (1996). A Hierarchical Linear Modeling Approach to Studying College Effects Corinna A. Ethington. In Smart, J. C. (Ed.) Higher Education: Handbook of Theory and Research (Volume XII). New York: Agathon Press.

Feldman, K. A. (1989a). Instructional effectiveness of college teachers as judged by teachers themselves, current and former students, colleagues, administrators, and external (neutral) observers. Research in Higher Education, 30, 137-194.

Feldman, K. A. (1989b). Association between student ratings of specific instructional dimensions and student achievement: Refining and extending the synthesis of data from multisection validity studies. Research in Higher Education, 30, 583-645.

Feldman, K. A. (1997). Identifying exemplary teachers and teaching: Evidence from student ratings. In R. P. Perry & J. C. Smart (Eds.), Effective teaching in higher education: Research and practice (pp. 368-395). New York: Agathon

Feldman, K. A. (1998). Reflections on the study of effective college teaching and student ratings: One continuing quest and two unresolved issues. In J. C. Smart (Ed.). Higher Education: Handbook of theory and research, Vol. 13 (pp. 35-74). New York: Agathon

Greenwald, A. G., & Gillmore, G. M. (1997a). Grading leniency is a removable contaminant of student ratings. American Psychologist, 52, 1209-1217.

Gillmore, G.M., Kane, M.T., & Naccarato, R.W. (1978). The generalisability of student ratings of instruction: Estimates of teacher and course components. Journal of Educational Measurement, 15, 1-13.

Goldstein, H. (1995). Multilevel statistical models (2nd ed.). London: Arnold.

Goldstein, H., Rasbash, J., Plewis, I., Draper, D., Browne, W., Yang, M., Woodhouse, G., & Healy, M. (1998). A user's guide to MLwiN. London: Institute of Education, University of London.

Harmon, G., (1999). Vouchers or 'student centred funding': the 1996-98 Australian review of higher education financing and policy. Higher Education Policy, 12, 219-235.

Hill, P.W., & Rowe, K.J. (1996). Multilevel modeling in school effectiveness research. School Effectiveness and School Improvement, 7, 1-34.

Hill, P.W., & Rowe, K.J. (1998). Modeling student progress in studies of educational effectiveness. School Effectiveness and School Improvement, 9 (3), 310-333.

Hockey, J. (1995). Getting too close: A problem and possible solution in social science PhD supervision. British Journal of Guidance & Counselling, 2, 199-210.

Holdaway, E. A. (1996). Current issues in graduate education. Journal of Higher Education Policy & Management, 18 , 59-74.

Jöreskog, K.G., & Sörbom, D. (1999). LISREL 8.30. Chicago, IL: Scientific Software International Inc.

L'Hommedieu, R., Menges, R. J., & Brinko, K. T. (1990). Methodological explanations for the modest effects of feedback. Journal of Educational Psychology, 82, 232-241.

Marsh, H. W. (1984). Students' evaluations of university teaching: Dimensionality, reliability, validity, potential biases and utility. Journal of Educational Psychology, 76, 707-754.

Marsh, H.W. (1987). Students' evaluations of university teaching: Research findings, methodological issues, and directions for future research. International Journal of Educational Research, 11, 253-388. (Whole Issue).

Marsh, H. W., & Bailey, M. (1993). Multidimensionality of students' evaluations of teaching effectiveness: A profile analysis. Journal of Higher Education, 64, 1-18.

Marsh, H.W. & Dunkin, M. (1997). Students' evaluations of university teaching: A multidimensional perspective. In R.P. Perry & J.C. Smart (eds.), Effective Teaching in Higher education: Research and Practice. (pp. 241-320). New York: Agathon.

Marsh, H.W., & Roche, L.A. (1993). The use of students' evaluations and an individually structured intervention to enhance university teaching effectiveness. American Educational Research Journal, 30, 217-251.

Marsh, H.W., & Roche, L.A (1997). Making students' evaluations of teaching effectiveness effective. American Psychologist, 52, 1187-1197.

Paper Presented at the Joint AARE/
NZARE Conference, Auckland, 2003
Australian Association
for Research in Education

Marsh, H. W., & Roche, L. A. (2000). Effects of grading leniency and low workloads on students' evaluations of teaching: Popular myth, bias, validity or innocent bystanders? . Journal of Educational Psychology, 92,:202-228.

McKeachie, W. J. (1979). Student ratings of faculty: A reprise. Academe, 65, 384- 397.

McKeachie, W .J. (1997a). Good teaching makes a difference - And we know what it is. In R. P. Perry & J. C. Smart (Eds.), Effective teaching in higher education: Research and practice. (pp. 396-411). New York: Agathon.

McKeachie, W. J. (1997b). Student ratings: The validity of use. American Psychologist, 52, 1218-1225.

McKinnon, K. R., Walker, S. H., Davis, D. (2000). Benchmarking: A manual for Australian universities. Canberra: Australian Department of Education, Training and Youth Affairs.

Monk, D.H. (1992). Education productivity research: An update and assessment of its role in education finance reform. Education Evaluation and Policy Analysis, 14, 307-332.

Pearson, M. (1996). Professionalising Ph.D. education to enhance the quality of the student experience. Higher Education, 32, 303-320.

Ramsden, P. (1991). A performance indicator of teaching quality in higher education: The Course Experience Questionnaire. Studies in Higher Education, 16, 129-150.

Rowe, K.J. (1999). Multilevel structural equation modeling with MLn/MLwiN & LISREL 8.30: An integrated course (3rd ed.) The 6th ACSPRI Winter Program in Social Research Methods and Research Technology, The University of Western Australia. Centre for Applied Educational Research, The University of Melbourne.

Rowe, K.J., & Hill, P.W. (1998). Modeling educational effectiveness in classrooms: The use of multilevel structural equations to model students' progress. Educational Research and Evaluation, 4, 307-347.

Rowe, K.J., Hill, P.W, & Holmes-Smith P. (1995). Methodological issues in educational performance and school effectiveness research: A discussion with worked examples (Leading article). Australian Journal of Education, 39, 217-248.

Rowe, K.J., & Rowe, K.S. (1999). Investigating the relationship between students' attentive-inattentive behaviors in the classroom and their literacy progress. International Journal of Educational Research, 31 (1/2 – Whole Issue), 1-138.

Rowe, K.J., Turner, R., & Lane, K. (in press). The 'myth' of school effectiveness: Locating and estimating the magnitudes of major sources of variation in students' Year 12 achievements within and between schools over five years. School Effectiveness and School Improvement.

Scheerens, J., & Bosker, R. (1997). The foundations of educational effectiveness. Oxford: Pergamon.

Paper Presented at the Joint AARE/
NZARE Conference, Auckland, 2003
Australian Association
for Research in Education

Footnotes

(1) In Australia, research higher degree students complete an extensive thesis at either the PhD level or the Masters Honors level. In each case, the focus of the research training program is primarily the final thesis that is externally examined by two (Masters Honors) or three (PhD) examiners who are external to the university and will often be from universities outside of Australia. The progam typically involves little or no formal coursework or qualifying examinations, although there typically are annual reports completed by students and supervisors to monitor progress. For purposes of the present investigation, we use the term PhD students generically to include all research higher degree students and the acronym PhD students' evaluations to refer to evaluations of the supervision and research training experience by these students.

(2) It is curious that most of the differences between universities involve a single NZ university where the ratings were exceptionally low and the number of students was exceptionally large (representing more than 15% of the students from the entire study and more than double the number of research students from the largest Australian university). Although clearly beyond the scope of the present investigation, it appears that this apparently anomalous result represents idiosyncratic circumstances (e.g., the sample may have included coursework masters students who did not receive much in the way of research supervision because they were not enrolled in a research degree course) at a single NZ university. Although unexplained, the anomalous results from one of the 32 universities do not provide an adequate basis for the claim that PREQ responses were able to discriminate among universities, particularly given that there are no significant differences between the other 31 universities and that the variance component representing differences across all universities is not statistically significant.