Journal of Educational Psychology
2011, Vol. 103, No. 3, 733–748

0022-0663/11/$12.00 DOI: 10.1037/a0024221

Use of Student Ratings to Benchmark Universities: Multilevel Modeling of Responses to the Australian Course Experience Questionnaire (CEQ)

Herbert W. Marsh
University of Oxford and University of Western Sydney

Paul Ginns
University of Sydney

Alexandre J. S. Morin
University of Sherbrooke and University of Western Sydney

Benjamin Nagengast
University of Oxford

Andrew J. Martin
University of Sydney

Recently graduated university students from all Australian Universities rate their overall departmental and university experiences (DUEs), and their responses (N = 44,932, 41 institutions) are used by the government to benchmark departments and universities. We evaluate this DUE strategy of rating overall departments and universities rather than individual teachers, and we juxtapose it with the traditional use of student ratings to evaluate individual teachers (SETs). Multilevel analyses of DUE overall ratings were not able to discriminate well between universities or departments—few universities or departments differed significantly from the grand mean. Although the a priori 5-factor structure for this DUE instrument was reasonably well-defined at the individual student level, none of the 5 factors separately or in combination discriminated well between departments or universities. In contrast to this pattern of results, we review studies showing that SETs do reliably differentiate between teachers and are valid in relation to many criteria of effective teaching. However, casual reviews of these research literatures should not use this support for SETs to justify the use of DUE-type strategies. We conclude that DUE-type ratings should be used with great caution, if at all, and should not be seen as an alternative to SETs.

Keywords: university educational experience, multilevel analyses, students' evaluations, benchmarking universities

Supplemental materials: http://dx.doi.org/10.1037/a0024221.supp

Students' evaluations of educational effectiveness are widely used to evaluate teaching effectiveness and educational quality in many countries across the world. These evaluations are important as feedback to students, teachers, departments, university administrators, governmental policymakers, and researchers. Hence, it is not surprising that substantive and methodological studies in this area have resulted in a huge research literature, one of the most widely studied topics in educational psychology.

The vast majority of this research is based on the traditional approach to students' evaluations of teaching, in which students in a specific class taught by a specific teacher evaluate the teaching effectiveness of their teacher, typically near the end of the term.[1] Ratings by all students within the class are aggregated to form class-average ratings that are used as feedback to teachers to improve their effectiveness; the ratings are sometimes also used for personnel decisions by administrators, coursework selection by

This article was published Online First July 4, 2011.

Herbert W. Marsh, Department of Education, University of Oxford, Oxford, England; and Educational Excellence and Equity Research Center, University of Western Sydney, Sydney, New South Wales, Australia. Paul Ginns and Andrew J. Martin, Faculty of Education and Social Work, University of Sydney, Sydney, New South Wales, Australia. Alexandre J. S. Morin, Department of Psychology, University of Sherbrooke, Sherbrooke, Quebec, Canada; and Educational Excellence and Equity Research Center, University of Western Sydney. Benjamin Nagengast, Educational Excellence and Equity Research Center, University of Western Sydney.

Correspondence concerning this article should be addressed to Herbert W. Marsh, Department of Education, University of Oxford, 15 Norham Gardens, Oxford OX2 6PY, United Kingdom. E-mail: herb.marsh@education.ox.ac.uk

[1] For purposes of clarity, we use the following terminology: *Class* is a particular teacher teaching a particular group of students over a teaching term (e.g., Professor Jones teaching Psychology 101 to a group of 30 students across the fall semester), *department* is an academic department within a particular university (e.g., the psychology department within ABC University), and *discipline* is a generic term for the subject matter that is not specific to a particular university (e.g., psychology as a discipline) or an aggregate of departments devoted to the same discipline across all universities. We avoid use of the term *course* because it can ambiguously be used to refer to a particular class (as is typical in North America) or to a department or discipline (as is typical in Australia and the United Kingdom).

students, and research (Marsh, 2007b). Hereafter, we refer to this approach as SETs.

More recently, use of student ratings of their overall experience in relation to an entire department or university—hereafter referred to as departmental and university experiences (DUEs)—has increased. Increasingly, the purposes of DUEs are to rank different universities or departments within universities across all universities within a given country or educational system or to rank departments within particular universities or across universities. The intended uses of DUEs are to provide information for prospective students to use in selecting universities and departments through published summaries of the results, present diagnostic feedback to universities and departments that is useful in improving educational experiences, and help governments to reward excellence by benchmarking universities. There is considerable interest in the use of DUEs because they are much more cost effective than SETs (e.g., students need only complete one survey at the end of their educational experience rather than surveys in each of their classes every year), DUEs can be standardized across a large number of different universities and departments (SETs are typically at least somewhat idiosyncratic to different universities), and they can be centrally organized and mandated by governmental agencies for benchmarking universities.

Both SETs and DUEs are based on student ratings, but there are critical differences between the two approaches. In traditional SET research, the teacher in a particular class is the primary unit of analysis: Students rate a particular teacher teaching a particular class, and then ratings are aggregated to provide class-average ratings across students within the class for that teacher. Hence, critical research issues are (a) the extent to which the SET ratings reliably distinguish between individual teachers, (b) whether the ratings of a single teacher in a particular class are stable and generalizable in relation to other classes taught by the same teacher, and (c) if differences between teachers have construct validity in relation to other indicators of teaching effectiveness at the teacher level.

In DUE studies, the university or department is the primary unit of analysis; ratings are aggregated across students within a department or university. Here critical research questions are (a) the extent to which student ratings can reliably differentiate between different departments or universities and (b) whether these differences have construct validity in relation to other measures of educational effectiveness. This is a critical distinction, because SET research has given little attention to comparisons across universities, whereas DUE research gives little attention to individual teachers (and, typically, instructs students not to consider them). The juxtaposition between SET and DUE approaches is a central focus of the present investigation. Is it better to focus on the effectiveness of individual teachers or on the overall educational experience at the level of the department or university; or are the two approaches complementary?

In an overarching review of student rating instruments used to collect feedback about effectiveness in higher education, Richardson (2005) concluded

> It is clearly necessary that such a questionnaire should be motivated by research evidence about teaching, learning and assessment in higher education and that it should be assessed as a research tool. The only existing instruments that satisfy these requirements are the SEEQ [Student Evaluation of Educational Quality; Marsh, 1984, 1987] (for evaluating individual teachers and course units) and the CEQ [Course Experience Questionnaire; Ramsden, 1991a, 1991b] (for evaluating programmes). (p. 404)

Following Richardson's (2005) evaluation, we consider the SEEQ to be a prototypical SET instrument (designed to evaluate the teaching effectiveness of individual teachers) and the CEQ to be a prototypical DUE instrument (designed to evaluate university/department experiences). Our goal in the present investigation is to evaluate the CEQ instrument that is used in the Australian DUE program and to juxtapose these results with those from traditional SET research. The Australian program is one of the oldest of its kind and has served as a model for other DUE programs. Although we focus on the CEQ, it is important to emphasize that our real interest is in the evaluation of the DUE approach more generally.

We note that the history of SET research reflects a substantive-methodological synergy approach (Marsh & Hau, 2007), which addresses long-standing, complex, and unresolved substantive issues through the application of new and emerging quantitative research methodology. However, the corresponding DUE research literature is sparse, and current best practice analyses are not routinely applied (see Cheng & Marsh, 2010). Hence, not only do our findings have important theoretical and practical implications for monitoring and improving educational effectiveness, they also demonstrate evolving methodology to address these critical substantive issues. In pursuit of these aims, we begin with a selective review of SET research, juxtapose SET and DUE research, develop theoretical and methodological frameworks for the evaluation of DUEs, and then provide empirical results that are based on this framework.

An appropriate multilevel perspective is a critical aspect of the present investigation. In educational psychology and educational research more generally, the use of multilevel (or hierarchical) modeling is broadly acknowledged as the most appropriate approach for evaluating hierarchical data (e.g., Goldstein, 1995; Raudenbush & Bryk, 2002). Almost all educational data are multilevel: Students are nested within classes, classes are nested within departments, departments are nested within universities, and so forth. Students do not rate educational and teaching effectiveness in a vacuum; they do so in the context of particular classes, departments, and universities. As illustrated in previous DUE research (Cheng & Marsh, 2010; Marsh, Rowe, & Martin, 2002), in educational psychology research more generally (e.g., Marsh, 2008) and in the present investigation, research, policy questions, data, and statistical analyses that are appropriate at one level of analysis may be inappropriate or even misleading when evaluated at the wrong level of analysis. Thus, whenever the data have a multilevel structure, multilevel modeling should be the statistical technique of choice. Thus, for both SET and DUE approaches, it is inappropriate to conduct single-level analyses that focus on individual students as the sole unit of analysis. Construct validity, reliability, and usefulness of the scores cannot be evaluated only at the level of individual students.

SETs in Higher Education

In higher education, there is a long history of research and much debate into the appropriate use of SETs (e.g., d'Apollonia &

Abrami, 1997; Feldman, 1997, 1998; Greenwald & Gillmore, 1997; Marsh, 1980, 1982, 1984, 1987, 1991, 2007b; Marsh & Roche, 1997, 2000; McKeachie, 1997). Within the SET literature, there is consensus that the rating of one teacher in one class, averaged across responses by students in that class, is the appropriate unit of analysis rather than individual student ratings (e.g., Marsh, 1987, 2007b). Thus, support for the construct validity of SETs 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 class (for further discussion, see Gillmore, Kane, & Naccarato, 1978).

Effective teaching is a hypothetical construct for which there is no adequate single indicator. Hence, the validity of SETs 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, 1982, 1984, 1987, 2007b; Marsh & Dunkin, 1997; McKeachie, 1979, 1997) have consistently shown that with careful attention to measurement and theoretical issues, SETs are (a) multidimensional, reliable, and stable; (b) primarily a function of the instructor who teaches a class rather than the class that is taught; (c) relatively valid against a variety of indicators of effective teaching; (d) relatively unaffected by a variety of variables hypothesized as potential biases (e.g., expected class grades, class size, workload, and prior subject interest); and (e) demonstrably useful in improving teaching effectiveness when coupled with concrete enhancement strategies in specific areas that teachers target for improvement.

The correlation between responses by any two students in the same class (i.e., single-rater reliability or variance component at the individual student level; Marsh, 1987, 2007b) is typically in the .20s for different SEEQ items, but the reliability of the *class-average* response depends on the number of students rating the class: .95 for 50 students, .90 for 25 students, .74 for 10 students, and .60 for five students. This research demonstrates that students responding to SETs are very good at differentiating between teachers in terms of teaching effectiveness (Marsh, 1987, 2007b) so long as at least 10 students are in the class or several classes are taught by the same teacher, even if there are fewer than 10 students in each class. Indeed, SETs of the same teacher are highly stable over time when evaluated by successive cohorts of students. Thus, Marsh (2007a) found almost no systematic changes in students' ratings of 195 teachers over a 13-year period. Marsh and Bailey (1993) further demonstrated that each teacher had a characteristic profile on the different SET factors (e.g., high on organization and low on enthusiasm) that was distinct from the profiles of other teachers and generalized across classes over a 13-year period. Hence, not only is the overall level of SETs stable over time (i.e., a given teacher consistently gets high or low ratings) but even the profile of SETs (e.g., high in enthusiasm, but low in organization) is consistent over time.

Over time, the same teacher is likely to teach many different classes and the same class is likely to be taught by many different teachers. Hence, it is also important to determine how much of the SETs are a function of the teacher who teaches a class and of the class that is taught. In trying to separate the effects of the teacher and the class, Marsh (1987; Marsh & Dunkin, 1997) reported that the correlation between overall teacher ratings of different instructors teaching the same class (i.e., a class effect) was −.05, whereas correlations for the same instructor in a different class (.61) and in two different offerings of the same class (.72) were much larger. This suggests that SETs are a function of the person teaching the class rather than the particular class being taught. These results support the validity of SETs as a measure of teacher effectiveness but not as a measure of course quality that is independent of the teacher.

Although some research suggests 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 SET programs, ratings for a given class are normed in relation to similar classes (in terms of student composition, level, and discipline), implying that such differences are not seen as important. Although the existence of a well-developed SET program is an appropriate quality assurance indicator at the university level, there is little evidence that aggregating SETs to the level of departments or whole universities provides an appropriate measure of effectiveness at the department or university level. Indeed, research suggests that responses to typical instruments are not even very effective at differentiating between specific classes (independent of the teacher who teaches the class; Marsh, 2007b), let alone departments or whole universities.

DUEs

Universities throughout the world are undertaking benchmarking exercises, measuring performance on specific indictors on a common metric that allows systematic comparison (McKinnon, Walker, & Davis, 2000). This is done so universities can compare themselves with other universities on appropriate indices to establish their current levels of performance and to initiate self-improvement. Benchmarking exercises require a comprehensive set of benchmark indicators that focus on outcomes, measure functional effectiveness, are systematically developed so as to have good construct validity, and appropriately differentiate between universities (or departments within universities). Although such benchmarking exercises are often based on indicators of research productivity, the DUEs benchmark universities and departments in relation to student perceptions of their educational experiences. As noted earlier, the CEQ is a prototypical DUE instrument.

CEQ

The CEQ (see review by Richardson, 2005) is designed to assess student experience of their overall educational experience on the basis of responses to 23 items covering five domains (good teaching, clear goals and standards, appropriate workload, appropriate assessment, generic skills development) and one additional overall satisfaction item (also see the wording of CEQ items in Supplemental Table 1 of the supplemental materials). The theoretical basis of the CEQ came from Ramsden's (1991a, 1991b, 1992; also see Entwistle & Ramsden, 1983) research that contrasted surface (memorization and reproducing) and deep (meaning) elements of intellectual growth in relation to learning, teaching, and curriculum design in higher education.

The CEQ was derived from an earlier instrument (the Course Perceptions Questionnaire; Entwistle & Ramsden, 1983; Ramsden, 1979) that reflected how students' perceptions of the academic context—including experiences of teaching, assessment, and programs as integrated entities—relate to their approaches to learning. Initial item development was based on interviews with U.K. students from Lancaster University about their perceptions of the academic context, which suggested functional relationships between perceptions of the academic context and subsequent approaches to study (Entwistle & Ramsden, 1983; Ramsden, 1976). Analyses of responses from 2,208 participants across 66 U.K. university departments (Ramsden & Entwistle, 1981) revealed eight broad factors, including Good Teaching (e.g., clarity of explanations, appropriate pitching of materials, teaching staff enthusiasm, and helpfulness of staff), Openness With Students, Independence in Learning, Clear Goals and Standards, and Appropriate Workload. However, correlations of the Course Perceptions Questionnaire scale scores with students' reports of approaches to study were relatively small, a finding replicated by Parsons (1988).

These limitations motivated the subsequent development of the CEQ (Elphinstone, 1990) and a national trial of a revised, shortened, 30-item version of the CEQ to test its suitability as a performance indicator in the Australian context (Ramsden, 1991a, 1991b). The overarching aim of the CEQ was to produce "quantitative data which permit ordinal ranking of units in different institutions, within comparable subject areas, in terms of perceived teaching quality" (Ramsden, 1991a, pp. 132–133). Subsequently, in the benchmark exercise conducted by McKinnon, Walker, and Davis (2000), the CEQ was specifically recommended to benchmark departments and universities in relation to student experiences of teaching, goals and standards, assessment practices, workload, generic skills, and overall satisfaction (i.e., Benchmark 6.10). McKinnon et al. also noted that "Inter-university comparisons are most sensibly made across like fields of study and disciplines, or with universities that have broadly comparable profiles" (p. 86). On the basis of Ramsden's (1991a, 1991b) research, the Australian government mandated that the CEQ be completed by all graduates from all Australian universities within a few months of graduation.

Exploratory factor analysis (EFA) and confirmatory factor analyses (CFA) of responses to the CEQ items reported by the Graduate Careers Council of Australia (GCCA, 2002) consistently identified the five CEQ factors, in support of the a priori design of the CEQ and previous research. Although the GCCA (2002) report did not summarize the goodness of fit for the EFA, the fit for the CFA was only moderate in relation to current standards of a good fit (e.g., CFI = .88), apparently reflecting several nonzero cross-loadings identified in EFA. The GCCA suggested that the use of separate scales might be warranted for some specific purposes, but the major emphasis in summaries of CEQ responses has been on the CEQ overall satisfaction rating and, sometimes, the good teaching scale. Wilson, Lizzio, and Ramsden (1997) summarized correlations between the CEQ scores and approaches to study, satisfaction, self-reports of generic skills development, and academic achievement across multiple studies as evidence for the validity of the CEQ scales (for a recent review of CEQ scale correlates, see Watkins, 2001).

The GCCA (2002) report thus provides evidence of the reliability of the five CEQ factors at the level of the individual student. However, this report did not pursue more appropriate multilevel analyses to assess the extent to which CEQ responses were able to differentiate between universities that would support their use in benchmarking. Nevertheless, other research (Badhni & Aungles, 2002; Patrick, 2003) suggests that CEQ responses are not able to distinguish between universities. This suggests that the CEQ ratings are not able to differentiate between universities.

The GCCA reports have given more attention to the question of whether CEQ responses can differentiate between departments within universities. Thus GCCA (2003) noted that

> Graduates may find it difficult to condense their experiences of an entire course into the single response required for each item. In addition, if the results are averages of course experience there is the real possibility that the items will fail to discriminate between courses. Nevertheless, even though the result is a broad-brush measure, it does not appear to be a problem for respondents. (p. 21)

GCCA (2002) evaluated differences between 10 broad disciplines of study and differences between departments within universities in separate analyses for each discipline. Nevertheless, across the 10 disciplines, variance in the CEQ ratings explained by the differences between universities remained low and varied from 1.0% to 6.8% ($M = 3.2\%$). In addition, consistent with results based on DUE studies in the United Kingdom (Cheng & Marsh, 2010) and Australia (Marsh et al., 2002), the GCCA also found that there was little variance in the CEQ responses that could be explained by student background characteristics. Particularly given that these CEQ results were based on single-level analyses that ignored the multilevel structure of the data and are likely to bias the results in favor of showing greater differentiation, this issue warrants further consideration with more appropriate multilevel analyses.

As one of the first instruments to be used in a large-scale DUE program, the CEQ served as a model for similar undertakings in Australia (e.g., the Postgraduate Research Experience Questionnaire [PREQ]; Australian Council for Educational Research [ACER], 2000; also see Marsh et al., 2002) and DUEs developed in other countries (e.g., the National Student Survey [NSS] in the United Kingdom; Higher Education Funding Council of England, Quality Assurance Agency for Higher Education, Universities UK, and Standing Conference of Principals, 2001; Higher Education Funding Council of England, 2003; also see Barrie & Ginns, 2007). The intent of the CEQ, like that of the DUEs that followed it, was to provide an overall perspective of student experience at the level of the department or university, and the results of this exercise are broadly available, for example, through *The Good Universities Guide to Australian Universities* (http://www.gooduniguide.com.au/) used by potential students to select universities. Hence, an evaluation of CEQ in relation to its ability to differentiate between departments and universities is important in substantive and policy issues specific to the use of the CEQ, but it also has important implications for more general DUE strategies that have been implemented elsewhere.

PREQ: Unit of Analysis Issue

Critical concerns in the evaluation of the PREQ were (a) the ability of DUEs to differentiate between universities and departments and (b) the appropriate unit of analysis (ACER, 2000; Marsh et al., 2002). In part on the basis of the success of the CEQ, where the focus was primarily on undergraduate teaching programs, the Australian government commissioned the development and evaluation of the PREQ (ACER, 2000) to provide a measure of the experience of postgraduate research students (e.g., doctoral candidates and research masters) as part of a large-scale national benchmarking exercise of the postgraduate research programs of Australian universities. The PREQ was sent to all graduating research students to evaluate their research experiences.

The intent of the PREQ, like the other DUE instruments, was to compare whole universities and departments rather than individual supervisors (students were specifically instructed not to name their supervisor). On this basis, Marsh et al. (2002) argued that PREQ responses should minimally be able to differentiate between universities and departments. They found that PREQ responses had reasonable psychometric properties (factor structure and reliability) at the level of the individual student. However, multilevel statistical models showed that responses from students attending 32 Australian and New Zealand universities did not reliably differentiate between universities or departments. They concluded that PREQ responses should not be used to benchmark universities or departments. On the basis of this research, Richardson (2005) similarly concluded that the PREQ “ did not discriminate among different universities or among different disciplines at the same university” (p. 399), leading to “considerable skepticism about whether it provides an adequate basis for benchmarking universities or disciplines” (p. 399).

United Kingdom NSS

In the United Kingdom, the NSS is used to gather feedback from final year university students about their department and university experience. Begun in England, it has become a United Kingdom–wide initiative in which all publicly funded higher education institutions are mandated to participate annually (HEFCE, 2003). In 2001, HEFCE proposed a new method for evaluating quality assurance in teaching and learning in higher education and established a task group to further develop their objectives of promoting teaching and learning; to provide appropriate information to students, employers, and others about the quality at each higher education institution; and to provide each higher education institution information to use in their enhancement activities. Suggestions were provided to achieve a level of consistency across all U.K. universities by administering a national student questionnaire modeled on the Australian CEQ. The task group proposed and ran trial pilot versions in 2003 and 2004 before launching the NSS in 2005. It has since been administered annually and is intended to be used in the foreseeable future.

Cheng and Marsh (2010; also see Marsh & Cheng, 2008) evaluated the ability of the NSS overall rating to differentiate between universities and departments. On the basis of the nature of the NSS data and previous research (Marsh et al., 2002), they argued that the most appropriate analysis should be based on a multilevel model with three levels (students, departments, and universities). In their analyses of NSS responses they found that differences between universities were highly significant from a statistical perspective, primarily because of the very large sample sizes (*N*s = 171,290 and 157,342 students in 2005 and 2006, respectively). However, differences between universities explained only about 3% of the variance in individual student responses to the overall rating, and this estimate was further reduced (to about 2.5%) after controlling for discipline and student background characteristics. Hence, there was much more variation in responses by students within each university than there was between the different universities.

Differences between departments in the Cheng and Marsh (2010) study were more complex to interpret. Although more variance was explained by differences between departments than between universities, the number of students in different departments was necessarily much smaller. Hence, few departments significantly differed from the grand mean. Although Cheng and Marsh also evaluated effects of a variety of student background differences, these had little effect on overall ratings and almost no effect on the ability of NSS responses to differentiate between universities and departments. Cheng and Marsh thus recommended that NSS ratings should be used with caution, if at all, for comparisons between U.K. universities, departments within the same university, or departments across universities.

Present Investigation

Each year, approximately four months after completing a university coursework degree, all Australian university students were invited by their university and the GCCA to complete the CEQ. Participation in the data collection is mandated by the Australian government. Instruments are sent to all graduating students from all Australian universities—this is not a voluntary sample of convenience. Students are strongly encouraged to participate, and response rates (57% in the present investigation) are very good for mail-survey studies. Although the data collection is conducted separately by each institution according to standardized procedures, the actual analysis and reporting of the results is conducted centrally by the GCCA and the ACER (see GCCA, 2002).

The primary purpose of the CEQ is to provide student ratings of their educational experience across all Australian universities, and student responses are used for benchmarking departments and universities. A critical initial question that has not previously been addressed from an appropriate multilevel perspective is whether CEQ responses can reliably differentiate between universities or departments. If they are not able to do so, then their use in benchmarking exercises and publication in *The Good Universities Guide to Australian Universities* is questionable. Hence, the results of the present investigation have important implications for the appropriate use of the CEQ ratings and of the DUE approach more generally, associated public policy, and the research literature on educational quality assurance.

In the present investigation, we initially focus on the overall rating item as a single best indicator of students’ overall educational experience. Although it might also be appropriate to use a simple average response to all CEQ items for this purpose, this approach implicitly assumes that each of the specific CEQ factors (or items) is equally important. Also, the use of this unweighted average further assumes that there are no additional aspects of satisfaction with educational experience beyond those that have been measured by the CEQ. Clearly, each of these implicit assumptions is problematic and requires further consideration (see related discussion by Cheng & Marsh, 2010). In support of this rationale based on the NSS (Cheng & Marsh, 2010), higher order factor analyses of NSS responses indicated that there was a single higher order factor, that the factor loadings of first-order factors on the higher order factor were quite varied, and that the strongest relation was with the single overall rating. In supplemental analyses based on CEQ ratings (see Appendix 1 in the supplemental materials), we also found that there was a single higher order factor and that the highest factor loading was for the CEQ overall rating item. Furthermore, the CEQ's overall rating item is also the primary source of information used in ranking universities and departments in policy-related applications. Although Cheng and Marsh argued that the overall rating item provided the best overall summary, they also conceded that ratings of specific factors might provide more useful information than is available from the overall rating item. Following from this earlier DUE research, we also begin with an initial focus on the CEQ overall rating item but then evaluate in more detail whether results based on the overall rating item generalize to the specific CEQ factors, whether some specific CEQ factors are better at differentiating between universities or departments than the overall ratings, and whether some universities or departments have a distinctive profile of CEQ factors.

Method

Participants

In 2001, the CEQ was completed by recently graduated undergraduate (bachelor degree) students from all Australian universities (N = 44,932 students, 41 institutions, 57% response rate). Excluded were students who completed the CEQ items in relation to a postgraduate degree or who did not complete the CEQ items (the CEQ was part of a larger package of materials and some students did not complete the CEQ section). Analysis of disciplines (across universities) and departments (within universities) is based on three alternative classification schemes used by the GCCA to summarize the results that differ in terms of the number of categories (10, 43, or 186 categories). A majority of the respondents were female (59%). Although respondents varied from 19 to 81 years of age, most (69%) were less than 25 years old (mean age = 26.2 years). In addition to CEQ ratings, students completed a questionnaire regarding background-demographic characteristics. Although these are potentially important in their own right, our primary focus is the extent to which these student background differences contributed to apparent differences between universities and departments. The set of 12 background variables were employment (the presence or absence of paid work in the final year of school), gender, age, Aboriginal or Torres Strait Islander, non-English-speaking background, presence of disability, highest level of previous education, mode of study (residential or external), mode of attendance (part time or full time), mode of fee payment (fee deferred, Australian fee paying, overseas fee paying, other), mobility (changed city or changed state to attend university), and country of permanent residence (for further discussion of these variables and their effects, see GCCA, 2002).

Instruments

The CEQ is designed to assess students' educational experience on the basis of responses to 23 items covering five domains (good teaching, clear goals and standards, appropriate workload, appropriate assessment, and generic skills development) and one additional overall satisfaction item ("Overall, I am satisfied with the quality of this course"). Graduates are asked to think about their educational program as a whole rather than individual classes or teachers. They are also assured that their individual responses are confidential, but the database of anonymized responses from all students is made publicly available. Graduates responded to each question on a 5-point Likert-type scale ranging from 1 (*strongly disagree*) to 5 (*strongly agree*). The CEQ was administered to graduates 4 months after their degree completion (April for the majority who completed their degree at the end of the calendar year and October for a smaller proportion of students who completed their degree midyear).

Statistical Analyses

Statistical analyses were conducted with the MLwiN 2.02 statistical package (Rasbash, Steele, Browne, & Prosser, 2005). Consistent with the design and intent of the CEQ and consistent with previous DUE studies (e.g., Cheng & Marsh, 2010; Marsh et al., 2002), we evaluated a multilevel model with three levels (L1 = students, L2 = departments, L3 = university) that included fixed effects to control for student background characteristics and differences between disciplines that generalize across universities. In preliminary factor analyses (presented in detail in Appendix 1 of the supplemental materials), we evaluated a priori five-factor structural model of CEQ responses. We used CFA and newly developed exploratory structural equation modeling (ESEM) techniques that integrate many of the best features of confirmatory and exploratory approaches to factor analysis (Marsh et al., 2009, 2010). A comparison of the two approaches (see Appendix A) shows that both analyses identified the five a priori factors. However, the ESEM fit was much better than the CFA fit, reflecting the finding that several CEQ items had moderate cross-loadings that were constrained to be zero in the CFA. On the basis of these preliminary analyses, for purposes of the present investigation, the five CEQ specific factors are represented as ESEM factor scores (produced by the Mplus statistical package; Muthén & Muthén, 2010).

For large-scale studies, the inevitable missing data represents a potentially important problem, particularly when the amount of missing data exceeds 5% (e.g., Graham, 2009; Graham & Hofer, 2000). Here, however, the amount of missing data was extremely small (0.3% for the overall rating and only 0.8% across all CEQ items), so that it was unlikely to be a serious problem. Nevertheless, a growing body of research has emphasized potential problems with traditional pairwise, listwise, and mean substitution approaches to missing data (e.g., Graham, 2009; Graham & Hofer, 2000), leading us to implement the expectation maximization algorithm, a widely recommended approach to imputation for missing data, as operationalized using missing value analysis in SPSS (Version 17). For present purposes, all student background variables were categorical and represented by a set of dummy variables. For these categorical variables, we added an additional category for missing data (see Cheng & Marsh, 2010). Thus, for example, gender was categorized into three categories (1 = male, 2 = female, 3 = missing) and represented by two (1 degree of freedom) contrasts: (a) males versus females and (b) missing versus males and females. In this way, we did not need to impute missing values for categorical variables and we were able to assess whether students with missing data on each categorical variable differed from those who did not.

Results and Discussion

Do Global Satisfaction Ratings Differentiate Between Universities and Departments?

We evaluate this overarching question in relation to variance components associated with departments and universities and with a graphic representation of these differences (*caterpillar plots*). Both variance components and the corresponding caterpillar plots are based on a set of 12 three-level models (wherein Level 1 = students, Level 2 = departments, and Level 3 = university) consistent with the hierarchical structure of the CEQ data (see Table 1). The models differ in terms of the number of department classifications included as random effects (10, 43, or 186 classifications), the number of discipline classifications included as fixed effects (0, 10, 43, or 186 classifications), and the inclusion or noninclusion of the set of 12 student characteristics as fixed effects.

Differentiation between universities. We begin with the variance components associated with universities for the models that do not control for discipline or student characteristics (Model 1; see Table 1) but vary in the number of department classifications considered within universities: 10 (Model 1A), 43 (Model 1B), or 186 (Model 1C). Although there is some variation in the results depending on the model, the variance components for differences between universities are all close to .01 (.009 in Model 1A, .013 in Model 1B, .014 in Model 1C). Thus, only about 1% of the variance in CEQ responses is explained by differences between universities.

What does it really mean that differences between universities explain only 1% of the variance in CEQ ratings (i.e., variance components of .01)? The results are clearly significant from a statistical perspective, but are even these small differences substantively meaningful? It might be possible to posit subjective classifications of variance components with labels such as *trivial*, *small*, *medium*, or *large*. However, these subjective labels would be highly idiosyncratic to the particular application and actually provide less information than the variance-explained metric that is more objective. Instead, we use caterpillar plots (see Figure 1, based on a selection of models already considered in Table 1) to provide a particularly useful graphical approach to evaluating substantive meaningfulness in multilevel data.

Consider initially the first caterpillar plot for universities (Model 1A for universities in Figure 1). The universities are ranked from left to right in terms of CEQ global satisfaction rating. For each university, the university estimate and an error bar (95% confidence interval) are presented. The sizes of these error bars reflect unreliability in the ratings: They are smaller (indicating less error) when the agreement among students within a given university is better and when the number of students responding from that university is larger. Because these are residuals based on standardized (M = 0, SD = 1) values, the differences are in terms of standard deviation units with a mean of zero across all universities (the dashed line in Figure 1). If a university has an error bar completely above or completely below the mean of all universities (i.e., the dashed line), it is significantly above or below average in a statistical sense. If the error bar contains the grand mean (i.e.,

Table 1
Residual Variance (RV) Associated With University, Department, and Student Levels: Overall Satisfaction Rating

	Null Model 1		Background Model 2		Discipline Model 3		Discipline and background Model 4	
Level of analysis	*RV*	*SE*	*RV*	*SE*	*RV*	*SE*	*RV*	*SE*
Model A, 10 disciplines								
University	.009	.003	.004	.002	.009	.003	.004	.002
Department	.033	.004	.037	.004	.021	.003	.024	.004
Student	.967	.006	.961	.006	.967	.006	.962	.006
Model B, 43 disciplines								
University	.013	.004	.009	.003	.011	.003	.007	.002
Department	.057	.004	.058	.004	.037	.003	.036	.003
Student	.943	.006	.938	.006	.943	.006	.940	.006
Model C, 186 disciplines								
University	.014	.004	.010	.003	.010	.003	.007	.002
Department	.066	.004	.065	.004	.032	.003	.032	.003
Student	.933	.006	.928	.006	.932	.006	.929	.006

Note. A series of multilevel models with three levels (Level 1 = universities, Level 2 = departments within universities, Level 3 = students) were conducted on the overall satisfaction rating as the outcome variable, Models vary in terms of the fixed effects. Model 1 is a variance component null model with no fixed effects. Models 2–4 include background (i.e., student demographic characteristics discussed earlier; Model 2), discipline (Model 3), or both background and discipline (Model 4) variables as fixed effects. For each of these models, the classification of departments and disciplines varied such that the number of disciplines was 10 (Model A), 43 (Model B), and 186 (Model C).

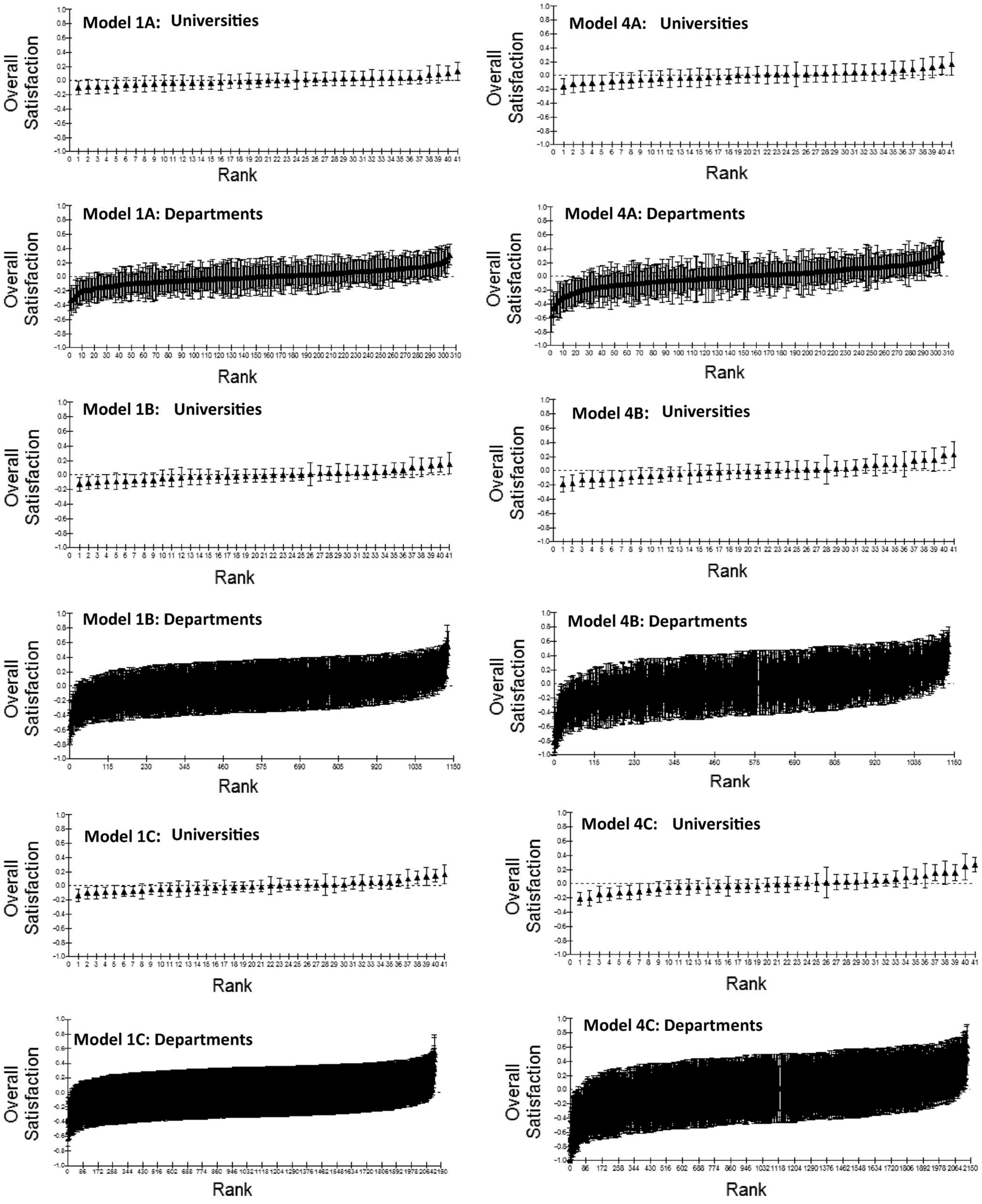

Figure 1. Selected caterpillar plots showing differentiation between departments and universities for models with no fixed effects (Model 1) and for models with discipline and student characteristics as fixed effects (Model 4). Also see Table 1.

overlaps the dashed line), then the university is not significantly different from the mean of all universities. For Model 1A, the 95% confidence intervals for almost all the universities contain the overall mean.

Some of the variation attributed to universities is likely to be due to discipline differences that generalize across universities (i.e., different universities have a different mix of disciplines) as well as differences in student characteristics. To explore these possibilities, we consider variance components associated with the same set of models in which we control student background characteristics (Model 2 in Table 1), disciplines (Model 3), or both discipline and student characteristics (Model 4). Variance components associated with differences in universities are systematically smaller in models that control for student characteristics and discipline fixed effects. However, because so little variance was explained by differences between universities to begin with, these reductions are not large. For example, the largest variance component is for Model 1C (.014) when no fixed effects are controlled. Controlling for either student characteristics (Model 2C) or discipline fixed effects (Model 3C) reduces the variance component to .010, whereas controlling for both student characteristics and disciplines further reduces the variance component to .007 (Model 4C). Across all of the caterpillar plots, the results are consistent. Only a few universities are significantly above average and only a few universities are significantly below average, but most universities do not differ significantly from the average across all universities. Furthermore, even if these small differences in the most extreme universities are significant from a statistical perspective (because of the large sample sizes), these differences are very small; they would not be significant with a more demanding test of statistical significance (e.g., 99% confidence intervals) or if the sample size was smaller, and they are apparently too small to be of practical significance (but see subsequent discussion in the Implications section). In summary, consistent with results based on the corresponding variance components (see Table 1), the caterpillar plots results show that CEQ responses are not very good at differentiating between universities—their primary purpose.

Differentiation between departments. Are CEQ responses able to differentiate between departments? Again, it is relevant to evaluate variation attributed to departments after controlling student background characteristics (Model 2 in Table 1), disciplines[2] (Model 3), or both discipline and student characteristics (Model 4). Controlling for student characteristics had almost no effect on variance components (i.e., comparison of Models 1 and 2 or of Models 3 and 4). However, the variance components associated with departments decreased systematically for models controlling for discipline fixed effects (Models 2 and 4). Thus, for the models that control disciplinary differences, the variance components were small (e.g., .032 in Models 3C and 4C). Although discipline fixed effects (i.e., discipline differences that generalize over universities) are of interest in their own right (for further discussion, see GCCA, 2002), our focus is on how their inclusion influences the variance components associated with departments. The results show that close to half of the variance due to differences between departments can be explained in terms of discipline effects that generalize across universities.

Next, we consider caterpillar plots for departments. These differ from the university caterpillar plots in several ways. It is important to note that differences between departments (i.e., the corresponding variance component) are larger than differences between universities. However, the error bars associated with departments are much wider (reflecting more error due primarily to the number of students in each department within a particular university necessarily being much smaller than the total number of students from the university). Hence, even though differences between departments are larger than differences between universities, the vast majority of departments do not differ significantly from the grand mean (i.e., the dashed line). Furthermore, although differences between departments increase as the number of classifications increases, the sizes of the error bars increase even more (because of the smaller numbers of students). Hence, differentiation between departments is not improved by use of a more detailed department classification.

Do CEQ Specific Factors Responses Reliably Differentiate Between Universities and Departments?

Results presented thus far are based on the global satisfaction ratings. Although clearly defensible in terms of the intended use of CEQ ratings and current practice that emphasizes the CEQ overall rating, the CEQ was also designed to measure five a priori factors based on 23 items. Hence, in this second set of analyses, we applied a subset of the models considered for the global satisfaction rating separately for each of the five specific CEQ factors. In particular, we are interested in whether evidence for differentiation among universities or departments is any greater for the specific factors than it is for the global satisfaction rating already considered in detail. We begin with separate analyses of each of the five CEQ factors—based on factor scores—and then consider the multivariate profiles of the factors.

On the basis of results reported earlier for the global satisfaction ratings, we chose a subset of three multilevel models and applied each of these models to the five specific CEQ factors. Model 1 is a simple three-level variance components model (wherein Level 1 = students, Level 2 = departments, and Level 3 = universities) with no fixed effects. Although the variance components for the university for the five CEQ factors (.002–.047, *Mdn* = .015; Table 2) are marginally larger than for the global ratings already discussed, they are still small, indicating little differentiation at the level of the university.

Variance components for the department are at least moderate in size for some CEQ factors. However, after controlling discipline and student characteristics, we found that variance components for the five specific CEQ factors in Model 4 (.013–.037; *Mdn* = .024) are comparable to that reported for the corresponding analysis of the CEQ global rating (see Table 1). Our interpretation of these results, consistent with earlier discussion and results with the

[2] Departments can be seen as discipline groups within each university. From this perspective, the variance components associated with departments represent the interaction between universities and disciplines—the extent to which differences in discipline vary as a function of universities. However, some of the variance in this interaction is confounded with main effects of discipline—the extent to which there are discipline differences that are consistent across universities. In this sense, it may not be appropriate to interpret the effects of the Discipline × University interaction (i.e., the effect of departments) without first controlling the main effect of discipline (i.e., those that generalize over universities) before.

Table 2
Residual Variance (RV) Associated With University, Department, and Student Levels: Selected Models for the Five Specific Factors

Level of analysis	Null model (discipline and background variables not controlled)						Discipline and background controlled					
	Model 1A		Model 1B		Model 1C		Model 4A		Model 4B		Model 4C	
	RV	*SE*	*RV*	*SE*	*RV*	*SE*	*RV*	*SE*	*RV*	*SE*	*RV*	*SE*
Teaching												
University	.047	.008	.048	.008	.046	.008	.048	.008	.046	.008	.043	.008
Department	.047	.003	.088	.003	.100	.003	.020	.003	.036	.003	.037	.003
Student	.942	.006	.909	.006	.893	.006	.895	.006	.895	.006	.880	.006
Goals												
University	.015	.003	.019	.003	.019	.003	.014	.003	.015	.003	.016	.003
Department	.031	.002	.051	.002	.060	.002	.015	.002	.025	.002	.023	.002
Student	.970	.006	.951	.006	.940	.006	.966	.006	.948	.006	.937	.006
Work												
University	.012	.003	.019	.003	.020	.003	.011	.003	.011	.003	.010	.003
Department	.086	.002	.121	.002	.120	.002	.018	.002	.034	.002	.028	.002
Student	.930	.006	.895	.006	.881	.006	.915	.006	.881	.006	.867	.006
Assessment												
University	.002	.002	.017	.002	.020	.002	.008	.002	.010	.002	.010	.002
Department	.096	.002	.147	.002	.164	.002	.017	.002	.030	.002	.026	.002
Student	.895	.006	.856	.006	.841	.006	.888	.006	.846	.006	.832	.006
Generic skills												
University	.005	.003	.006	.003	.006	.003	.005	.003	.005	.003	.006	.003
Department	.025	.002	.048	.002	.053	.002	.013	.002	.022	.002	.018	.002
Student	.974	.006	.954	.006	.946	.006	.968	.006	.950	.006	.941	.006

Note. A series of multilevel models with three levels (Level 1 = universities, Level 2 = departments, Level 3 = students) were conducted for each of the specific Course Experience Questionnaire factors. Models considered here are a subset of models used with the overall satisfaction rating (see Table 1).

global satisfaction rating, is that although there are significant variance components in specific CEQ factors associated with different departments, much of this variance generalizes across universities (i.e., is attributable to disciplines).

In supplemental analyses (see Appendix 2 in the supplemental materials), we also evaluated whether there are unique profiles of CEQ factors that are able to differentiate between universities or departments. However, this profile analysis indicated that neither universities nor departments are meaningfully differentiated in relation to either the average of the five CEQ factors (a level effect in profile analysis) or the unique profiles of CEQ factors (a shape effect in profile analysis).

In summary, conclusions based on separate analyses of each of the specific CEQ scales are largely consistent with those based on the global satisfaction rating. There is little evidence that any of the five specific CEQ factors are able to differentiate reliably between universities or departments—their primary purpose for which they are currently used by the Australian government and publications like the *The Good Universities Guide to Australian Universities*.

Summary, Implications, and Issues for Further Research

The minimum condition for evaluating the appropriateness of the CEQ responses for purposes of benchmarking universities is that they should be able to clearly discriminate between universities or departments. This is particularly important because there is surprisingly little appropriate research on this issue in the very large SET research literature (Marsh, 2007b) or even in the small but growing DUE research literature. Hence, our purpose in this article was to evaluate the ability of CEQ (and, more generally, DUE-type instruments) to differentiate between universities and departments. On the basis of Australian research of doctoral students' evaluations of their research experience, Marsh et al. (2002) argued that the appropriate unit of analysis for benchmarking exercises should be either universities or departments. Cheng and Marsh (2010) offered a similar argument for NSS ratings by U.K. students. Following this previous research, we argued that the most appropriate analysis should be based on a three-level multilevel model (wherein Level 1 = students, Level 2 = departments, and Level 3 = university).

Variance components at the university level were highly significant from a statistical perspective, primarily because of the very large sample sizes in nearly all universities. However, differences between universities explained very little variance in individual students' responses, and this estimate was further reduced when controlling the effects of discipline and student characteristics. Consistent with these results, caterpillar plots showed that very few universities had ratings that differed significantly from the grand mean. The results were consistent across the overall rating item, each of the five specific CEQ factors considered separately, and unique combinations of the specific CEQ factors.

Variance components at the level of departments are more complex to interpret. Not surprisingly, more variance is explained by differences among departments than by differences between universities. Within each university, there are systematic differences in the levels of satisfaction for students in different departments. However, inspection of caterpillar plots revealed that the error bars associated with departments were also much larger than those based on universities (because the numbers of students within departments is smaller than than the number of students within universities). Hence, very few departments differed significantly from the grand mean. Similarly, variance components for departments were larger when more detailed classifications were considered (i.e., 41 or 186 categories rather than 10 categories). However, reliable differentiation between departments requires not only large differences between groups but also a sufficiently large number of students within each group so that these differences are reliable. Particularly for the more detailed classifications, the number of students within each department is too small to reliably differentiate between departments. In summary, meaningful differentiation among departments is no better, and may perhaps be worse, than for universities. Again, the results were consistent across the overall rating item, the five specific CEQ factors, and unique combinations of the specific CEQ factors.

What do our results have to say about the DUE strategy more generally? Are our findings likely to be idiosyncratic to the CEQ instrument or the Australian setting, or are they likely to generalize to other DUE-type instruments? Because the CEQ program is the oldest DUE-type program and one basis of many other DUE programs, these results are important. Furthermore, our results replicate and extend findings based on entirely different instruments used among Australian postgraduate students (the PREQ; Marsh et al., 2002) and U.K. undergraduate students (the NSS; Cheng & Marsh, 2010). In summary, a growing body of research based on appropriate multilevel models calls into question the ability of DUE-like instruments to meaningfully differentiate between universities or departments.

An important overarching theme of the present investigation is that DUE-type and SET-type strategies are very different approaches to evaluating educational effectiveness. Furthermore, they have distinct research literatures that have quite different implications. Although the approaches might be complementary, there has been surprising little attempt to juxtapose the two approaches. Hence, it is important that casual reviews of these two research literatures do not fall into the trap of using support for SETs to justify the use of DUE-type strategies or the nonsupport of DUEs to oppose SETs.

Juxtaposition of SETs and DUEs

In an overarching review of student feedback instruments on teaching effectiveness in higher education, Richardson (2005) distinguished between SET-type instruments for evaluating individual teachers and DUE-type instruments like the CEQ for evaluating entire programs, noting that

> it is clearly sensible to seek feedback at a level that is appropriate to one's basic goals. If the aim is to assess or improve the quality of particular teachers, they should be the subject of feedback. If the aim is to assess or improve the quality of particular programs, then the latter should be the subject of feedback. Logically, there is no reason to think that obtaining feedback at one level would be effective in monitoring or improving quality at some other level (nor any research evidence to support this idea, either). (p. 401)

Nevertheless, Richardson (2005) went on to pose the question "Would a single questionnaire be suitable for all students?" (p. 401). In addressing this question in this final section, we briefly juxtapose SET and DUE research.

On the basis of reviews of the SEEQ responses and other well-designed SET-type instruments, student ratings of individual teachers teaching a specific class in which the students are enrolled are good at reflecting differences between individual teachers; they provide highly reliable, consistent, and stable differentiation between teachers (e.g., Marsh, 2007b). Furthermore, there is good evidence that SETs are valid in relation to a variety of indicators of effective teaching (student learning based on the multisection validity paradigm, teacher self-evaluations of their own teaching, observations of trained external observers, and the retrospective ratings of former students). They are also useful for improving teaching, personnel decisions, and student choice of teachers. In a study of the long-term stability of SEEQ responses, Marsh et al. (2007a) evaluated responses by a group of 195 teachers who had been evaluated continuously over 13 years (an average of 2.5 classes per year). He found that SEEQ responses systematically discriminated between good and poor teachers and that these differences were highly consistent over an extended period of time. A multilevel caterpillar plot based on these data (see Figure 2) shows that most teachers are consistently above average or con-

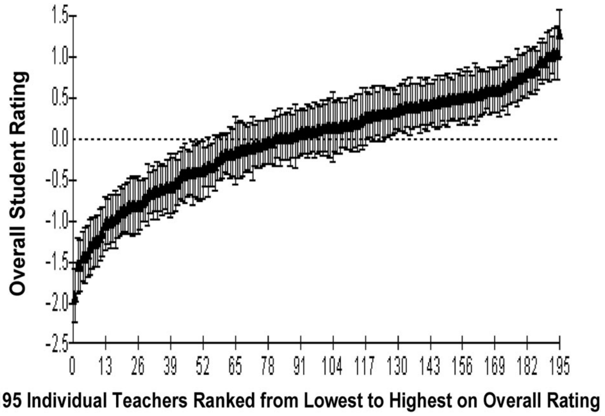

Figure 2. One hundred ninety-five individual teachers ranked from lowest to highest on overall rating of teaching effectiveness. Each teacher was evaluated continuously over 13 years (an average of more than 30 sets of ratings per teacher). Marsh (2007a) used multilevel growth modeling to show that ratings were highly stable over time. Here, we present a caterpillar plot based on a multilevel analysis to show that the mean ratings for the teachers (the 195 triangles) are highly differentiated and that the 95% confidence intervals are narrow relative to differences between teachers. Over the 13-year period, most teachers were consistently above average, below average, or close to average. These results are in contrast to caterpillar plots of Course Experience Questionnaire ratings (see Figure 1) showing little discrimination in relation to universities and courses of study within universities.

sistently below average. This differentiation between good and poor teachers is in stark contrast to the almost complete lack of differentiation between universities and departments based on CEQ responses (see Figure 1).

SETs are apparently not very good at benchmarking universities and departments. Indeed, existing research (Marsh & Bailey, 1993) suggests that they are not very good at differentiating even between classes independent of the teachers who teach the classes. Implications of these results are a double-edged sword. Because the intent of SETs is to measure the teaching effectiveness of individual teachers, these results highlight an important strength of SETs. Indeed, if SET responses were substantially confounded by the class that is being taught rather than the teacher who teaches the class, it would severely complicate the interpretation of the results in terms of the teaching effectiveness of individual teachers. However, this same characteristic reflects a weakness in terms of evaluating classes independent of the teachers who teach them, departments, or universities. This apparent limitation with SETs is one basis for proposing DUE-type instruments such as the CEQ. However, the results of the present investigation add to the existing research demonstrating that DUEs are not able to adequately differentiate universities or departments. On this basis, it seems that even this minimal basis of support for DUEs is lacking.

The results of the present investigation lead us to conclude that there is no support at all for the use of DUEs instead of SETs as a basis of diagnostic feedback to individual teachers or associated personnel decisions. Although SETs are also apparently ineffective in discriminating between universities and departments, there is little support for DUEs in relation to this purpose. Indeed, there is no evidence that DUEs are any better than SETs at discriminating between universities and departments.

Construct Validity of DUE-Type Responses: The Big Picture

Similar to the present investigation, DUE studies have focused on the ability of DUE-type responses to differentiate universities and departments. These studies clearly fall near the reliability end of the reliability–validity continuum. Reliable differentiation requires a combination of good agreement among students within each department or university and large numbers of students within each department or university.[3] However, the real construct validation issue is whether this differentiation is meaningfully related to other indicators of educational effectiveness. Reliable differentiation is not useful unless the differences can be interpreted meaningfully within a nomological network of other constructs.

Apparently, there is almost no appropriate research that even considers the construct validation of DUE-type ratings in relation to external criteria of effectiveness at the level of the university or department. What is needed is a systematic evaluation of the construct validity of DUEs, evaluating support for their convergent and discriminant validity in relation to external validity criteria (along the lines of the construct validity approach used to validate SETs; Marsh, 2007b). A limited amount of research suggests that students who give higher DUE ratings differ from other students on self-perceptions of student learning (e.g., Lizzio, Wilson, & Simons, 2002; Ramsden, 1991a, 1991b). However, as has been clearly demonstrated in SET research, findings at the level of individual students are largely irrelevant to construct validation at the level of teachers, departments, or universities—particularly when based on self-report responses by the same students. Thus, for example, Marsh et al. (2002) argued that aggregate university measures of PREQ ratings by research-degree students should be validated in relation to external criteria such as standardized measures of university research performance, number of Australian postgraduate research fellowships, attrition rates in doctoral programs, and related criteria such as those used by the Australian government to fund doctoral research programs. However, none of these external criteria were significantly related to PREQ ratings. Marsh et al. (2002) argued that given the lack of reliability of the PREQ ratings in relation to discriminating between universities, it is hardly surprising that PREQ ratings also lack validity in relation to these or other meaningful external criteria at the university level. Although we do not argue that these findings invalidate DUEs ratings of coursework degrees, the limited ability of CEQ responses to reliably discriminate between universities and departments does not augur well for more demanding tests of their construct validity in relation to external criteria.

Alternative Interpretations of CEQ Results

Lack of differentiation. Our results suggest that DUE-type responses are not very good at differentiating between universities. This pattern of results is consistent across three studies: the Australian PREQ (Marsh et al., 2002), the UK NSS (Cheng & Marsh, 2010), and the CEQ in the present investigation. We interpreted this as a shortcoming in the usefulness of the DUE approach—particularly for benchmarking universities. However, an alternative perspective might be that there really are almost no differences between universities and departments in terms of student experience, and this true lack of differentiation is accurately reflected in the DUE responses. Indeed, one could argue that in an ideal national system of universities with a focus on equity, there should be little or no difference between universities. Thus, the observed lack of differentiation should be a cause for celebration.

However, several arguments undermine this interpretation. Given that there are clear differences between universities and departments in relation to many indices of quality, this interpretation is implausible. Indeed, in a Higher Education Policy Institute report for the United Kingdom, Brown (2010) noted that

> Given the extraordinarily high previous educational attainment of students attending, say, Oxford or Cambridge, the substantially greater resources devoted to them, the greater intensity of study that they undergo, and other factors, it would in fact be a surprise if the outcomes of students from those universities were no higher than those of students from other universities who have far lower prior attainment, resources devoted to them, and so on. (p. 10)

Similarly, the Australian PREQ study showed that there were systematic differences between universities on objective indicators of quality that were not reflected in the PREQ ratings.

[3] This multilevel perspective on reliability at the organizational level is analogous to that based on reliability estimates at the item level for multiple items designed to measure the same construct; good reliability is a function of good agreement among items in the same group of items (rather than among students from a specific group) and a sufficiently large number of items (rather than students).

In contrast to the DUE strategy, we note that within the SET literature based on ratings of individual teachers, there is good evidence that student ratings do differentiate between good and poor teachers. Furthermore, this differentiation between teachers is stable, generalizable, and valid in relation to many indicators of good teaching. Why is there such a large difference between the two strategies? Students have been exposed to a wide variety of teachers, some good and some bad. Hence, they have a reasonable frame of reference against which to evaluate individual teachers. In contrast, students have typically experienced only a single department within a single university. Hence, they have no frame of reference against which to evaluate their experience with their department and university. Indeed, if the purpose of a quality control exercise is to compare different universities, it is illogical to base these comparisons on the judgments of persons who have no experience of different universities.

We also note that that this counterinterpretation would be stronger if there were relatively good agreement among students within departments and universities. However, this is not the case. There is considerable variation among DUE ratings at the level of the individual student, but these differences are not explained by the student's particular department (or by a wide variety of different student background characteristics). So what is the meaning of these large differences in DUE ratings at the individual student level? In the statistical models used here, these differences are interpreted as random error. Although it is always possible that additional research will support the construct validity of DUE ratings—and thus justify the interpretation of small differences as a desirable outcome—the onus for such evidence lies with those who advocate their usefulness. We argue that this explanation is improbable.

Finally, we note that even if this counterinterpretation were valid, it would not support the current use of DUE ratings to benchmark universities and departments, to differentially reward universities, or to provide potential students with information in the selection of universities.

Identification of outliers. Even if the DUE-type ratings are not very effective at differentiating between most universities and departments, perhaps they are useful in identifying outliers—the very best to be rewarded and the very worst to be improved (or eliminated). However, there are also problems with this application of DUEs.

Even if there were absolutely no differences between universities or departments, it would be possible to rank universities or departments such that there were a few highest and a few lowest. Particularly when the number of groups is large, there is an inevitable capitalization on chance—the highest and lowest ranked groups are so ranked in part due to good and bad luck. Regression to the mean implies that both highest and lowest ranked groups are really not so extreme and would not turn out to be so extreme when the exercise is repeated.

How large would differences have to be to be meaningful? Caterpillar plots of differences between departments and universities—even at the extreme—are very small. For example, even the most extreme universities are only about two tenths of a standard deviation above or below the grand mean. Furthermore, at least some of these even small differences are due to capitalization on chance. Although the differences at the level of the department are somewhat larger, the extent of capitalization on chance is also larger (i.e., error bars are larger and there are many more groups). Hence, at least from the perspective of a summative evaluation, the differences do not seem to be sufficiently large to be meaningful. By way of comparison, the caterpillar plot for SETs (see Figure 2) shows that there are many good teachers who are consistently (over 13 years) rated one standard deviation above the mean and many poor teachers who are rated one standard deviation below the mean. Hence, student ratings are able to differentiate educational effectiveness at the level of the teacher but not at the departmental or university levels.

Once again, we note that even if this counterinterpretation were valid, it would still not support the current use of DUE ratings to publicly benchmark universities and departments. Using DUE ratings for policy-based decisions regarding rewards and support might also mean keeping the rating confidential and limiting access to experts who are trained to reach a reasonable interpretation of these ratings.

Limitations, Unanswered Questions, and Directions for Further Research

The present investigation identified a host of methodological, substantive, and practical limitations with important policy implications. Critical questions raised here and in need of further research include the following:

1. Differences between universities explain only a very small amount of the variance in CEQ responses, and there is very little agreement among students within the same university. However, because of the very large number of students who responded from most universities, the results are highly significant from a purely statistical perspective. The question remains as to whether very small but statistically significant differences between universities are sufficiently large to help inform the choices of prospective students or other benchmarking activities based on CEQ responses.

2. Caterpillar plots (or related graphical strategies) apparently provide a useful strategy to interpret the practical meaningfulness of statistically significant variance components in DUE studies and multilevel studies more generally. Here the caterpillar plots provided an easily understood illustration of the poor ability of CEQ responses to differentiate between departments and universities (as well as the ability of SEEQ responses to differentiate between teachers). Caterpillar plots also showed why the larger department variance components associated with the use of more discipline categories did not translate into better differentiation between departments (i.e., the increase in error bars dues to smaller numbers of students offset the larger variance components). It is important to note that caterpillar plots are easily interpreted by applied researchers, policymakers, and members of the general public who might not have the technical expertise to understand variance components. Hence, we recommend that caterpillar plots should always be used to summarize the results of DUE studies, even those aimed at policymakers and the general public.

3. Department variance components are larger than university variance components. However, because the number of students in each department within a given university is so much smaller, few differ significantly from the grand mean. Differences between departments increase in size as the number of department categories increases, but the reliability of the means decreases (because the number of students is so much smaller). More research is

needed to achieve optimum balance between the number of discipline categories and the number of students in each department and to identify how well a single classification scheme can reflect the departments that actually exist at each university.

4. Although not a focus of the present investigation, it is important to evaluate potential biases in the CEQ results. The CEQ response rate was reasonably high for survey research, and results are not published in *The Good Universities Guide to Australian Universities* if the response rate for a given institution is below 50%. Nevertheless, further research is needed to evaluate more fully potential biases associated with nonresponse. Following a major study concerning the extent of bias, Graduate Careers Australia (2006) concluded that "non-response appears to introduce only low levels of bias into key GCA survey estimates" (p. 93). In the present study, the finding that the results are so little affected by the control of diverse student characteristics suggests that this might not be a serious problem. A more serious potential problem is the possibility that university staff could manipulate the ratings (for a discussion of such bias that represents a serious concern for the U.K. Higher Education Academy in relation to NSS ratings, see Cheng & Marsh, 2010). This is apparently more worrisome in Australia than the United Kingdom because each university has considerable input into the way the results are collected.

5. More research is needed to evaluate the stability of DUE ratings over time. Although GCCA reports routinely summarize the trends over time for the CEQ ratings averaged across all universities (e.g., GCCA, 2002), they do not evaluate the stability of ratings of specific universities and departments. Ratings are likely to be reasonably stable over a single year, in part because there is likely to be little change in the university, teachers, and even the student cohorts over such a short period. However, students using the CEQ to select universities typically will not graduate until four years later. The stability of the ratings is likely to decline with time—perhaps substantially—so that there is need for evaluations of stability over a longer period of time (see related research regarding the stability of rankings of schools across cohorts by Leckie & Goldstein, 2009).

However, tracking DUE ratings over time might provide formative evaluations that are useful in evaluating change, particularly if coupled with other indictors of educational effectiveness. Indeed, if an academic organizational unit instituted major curriculum changes, then instability in the ratings over time might be expected. Wilson et al. (1997), listing appropriate and inappropriate uses of the CEQ (see their Table 4, p. 49), emphasized intermittent planned use of the CEQ, especially for whole-program evaluation over time. A case study based on one department (Govendir, Ginns, Symons, & Tammen, 2009) provides some support for this use of DUE ratings. Although evidence is currently insufficient to evaluate this use of DUEs, it warrants further research.

6. Administrators typically compare CEQ ratings in different departments within their university as one way of evaluating which academic units are relatively stronger or weaker (e.g., Ginns, Marsh, Behnia, Cheng, & Scalas, 2009). However, the significant fixed effects associated with different disciplines suggest that there are systematic differences between disciplines averaged across all universities. Thus, it is possible for Department A to have higher ratings than Department B within a given university (implying that Department A is better than Department B), even though Department A has ratings below the average for the same department across all universities and Department B has ratings above the average for the same department across all universities (implying that Department B is better than Department A). Although we do not argue in favor of or against either of these apparently contradictory interpretations, we do note that administrators need further guidance in the interpretation of the ratings in relation to appropriate normative comparisons if they are expected to use the ratings to improve the quality of education at their universities.

Conclusions

In summary, we recommend that DUE ratings should only be used with extreme caution for benchmarking purposes. This caution applies to comparisons of ratings averaged across different universities and departments: either different departments within the same university or the same department across universities. Any such comparisons should be qualified in relation to interpretations of probable errors based on appropriate multilevel models—illustrated, for example, by caterpillar plots. These necessary cautions in the interpretation of DUE ratings also call into question their usefulness.

It is important to note that the conclusions in the present investigation replicate and extend similar conclusions by Marsh et al. (2002; also see Ginns et al., 2009) based on the Australian PREQ and by Cheng and Marsh (2010; also see Marsh & Cheng, 2008) based on the UK NSS. Although not a major focus of traditional SET research, SET studies also call into question the ability of ratings by students to differentiate classes independent of the teacher who teaches the class and, by implication, overall university or department experience. Taken together, the results indicate caution is needed when using DUE-type responses more generally. Although it might be premature to argue that DUE-type ratings are of little use in terms of providing appropriate feedback to students, employers, universities, and the general public, the onus is on advocates of these measures to demonstrate their construct validity in relation to their intended and actual use. A growing body of research suggests that they might not be very useful for their intended purposes. In considering the reliability, validity, and usefulness of student feedback in higher education, it is essential to distinguish between studies based on SETs and those based on DUEs.

Our take-home message is that policymakers should use extreme caution in the introduction of DUE-type programs and in the interpretation of DUE-type responses if such programs already exist. There is apparently no easy fix to the problems associated with DUEs. Indeed, it does not even make a lot of sense to base comparisons of different universities on the perceptions of students who mostly have experienced only a single university. Although future research with DUEs is clearly warranted, it is essential that such research is based on an appropriate multilevel perspective and seeks to validate interpretations of CEQ responses in relation to other indicators of educational effectiveness.

However, we also have a more positive take-home message. In contrast to DUEs, there is good support for the use of SETs in relation to the evaluation of individual teachers. Hence, it is important that systematic programs based on SETs are not replaced by DUE strategies. The apparent lack of support for

DUEs does not generalize to SETs and the good support for SETs does not generalize to DUEs. Thus, it is critical that reviews of the use of student ratings differentiate between these two dominant strategies.

References

Abrami, P. C., d'Apollonia, S., & Cohen, P. A. (1990). Validity of student ratings of instruction: What we know and what we do not. *Journal of Educational Psychology, 82,* 219–231. doi:10.1037/0022-0663.82.2.219

Australian Council for Educational Research. (2000). *Evaluation and validation of the trial Postgraduate Research Experience Questionnaires.* Camberwell, Victoria, Australia: Author.

Badhni, S., & Aungles, P. (2002, October). *The role of the Course Experience Questionnaire in quality assurance for the higher education sector.* Paper presented at the Seventh Quality in Higher Education International Seminar, Melbourne, Victoria, Australia.

Barrie, S., & Ginns, P. (2007). The linking of national teaching performance indicators to improvements in teaching and learning in classrooms. *Quality in Higher Education, 13,* 275–286. doi:10.1080/13538320701800175

Brown, R. (2010). *Comparability of degree standards?* (Higher Education Policy Institute report). Oxford, United Kingdom: Higher Education Policy Institute. Retrieved from http://www.hepi.ac.uk./files/47%20Comparability%20of%20degree%20standards.pdf

Cashin, W. E. (1988). *Student ratings of teaching: A summary of the research* (IDEA Paper No. 20). Retrieved from http://www.theideacenter.org/sites/default/files/Idea_Paper_20.pdf

Centra, J. A. (1993). *Reflective faculty evaluation.* San Francisco, CA: Jossey-Bass.

Cheng, J. H. S., & Marsh, H. W. (2010). National Student Survey: Are differences between universities and courses reliable and meaningful? *Oxford Review of Education, 36,* 693–712. doi:10.1080/03054985.2010.491179

Cohen, P. A. (1980). Effectiveness of student-rating feedback for improving college instruction: A meta-analysis. *Research in Higher Education, 13,* 321–341. doi:10.1007/BF00976252

d'Apollonia, S., & Abrami, P. C. (1997). Navigating student ratings of instruction. *American Psychologist, 52,* 1198–1208. doi:10.1037/0003-066X.52.11.1198

Elphinstone, L. (1990). *The development of the Course Experience Questionnaire* (Unpublished master's thesis). University of Melbourne, Melbourne, Victoria, Australia.

Entwistle, N. J., & Ramsden, P. (1983). *Understanding student learning.* London, United Kingdom: Croom Helm.

Feldman, K. A. (1989a). 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. doi:10.1007/BF00992392

Feldman, K. A. (1989b). 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. doi:10.1007/BF00992716

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, NY: Agathon Press.

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

Gillmore, G. M., Kane, M. T., & Naccarato, R. W. (1978). The generalizability of student ratings of instruction: Estimates of teacher and course components. *Journal of Educational Measurement, 15,* 1–13. doi:10.1111/j.1745-3984.1978.tb00051.x

Ginns, P., Marsh, H. W., Behnia, M., Cheng, J. H. S., & Scalas, L. F. (2009). Using postgraduate students' evaluations of research experience to benchmark departments and faculties: Issues and challenges. *British Journal of Educational Psychology, 79,* 577–598. doi:10.1348/978185408X394347

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

Govendir, M., Ginns, P., Symons, R., & Tammen, I. (2009). *Improving the research higher degree experience at the Faculty of Veterinary Science, The University of Sydney.* In *The student experience: Proceedings of 32nd HERDSA Annual Conference* (pp. 163–172). Milperra, New South Wales, Australia: Higher Education Research and Development Society of Australasia. Retrieved from http://www.herdsa.org.au/wp-content/uploads/conference/2009/papers/HERDSA2009_Govendir_M.pdf

Graduate Careers Australia. (2006). *Enhancing the GCA National Surveys: An examination of critical factors leading to enhancements in the instrument, methodology and process.* Canberra, Australian Capital Territory, Australia: Author.

Graduate Careers Council of Australia. (2002). *Course Experience Questionnaire 2001.* Canberra, Australian Capital Territory, Australia: Author.

Graduate Careers Council of Australia. (2003). *Course Experience Questionnaire 2002.* Canberra, Australian Capital Territory, Australia: Author.

Graham, J. W. (2009). Missing data analysis: Making it work in the real world. *Annual Review of Psychology, 60,* 549–576. doi:10.1146/annurev.psych.58.110405.085530

Graham, J. W., & Hofer, S. M. (2000). Multiple imputation in multivariate research. In T. D. Little, K. U. Schnable, & J. Baumert (Eds.), *Modeling longitudinal and multilevel data: Practical issues, applied approaches, and specific examples* (pp. 201–218). Mahwah, NJ: Erlbaum.

Greenwald, A. G., & Gillmore, G. M. (1997). No pain, no gain? The importance of measuring course workload in student ratings of instruction. *Journal of Educational Psychology, 89,* 743–751. doi:10.1037/0022-0663.89.4.743

Higher Education Funding Council for England. (2003). *National Student Survey: Administration* (Circular letter 22/2003). Retrieved from http://www.hefce.ac.uk/pubs/circlets/2003/cl22_03.htm

Higher Education Funding Council for England, Quality Assurance Agency for Higher Education, Universities UK, and Standing Conference of Principals. (2001). *Quality assurance in higher education: Proposals for consultation* (Consultation 01/45). Retrieved from http://www.hefce.ac.uk/pubs/hefce/2001/01_45/01_45.pdf

Leckie, G., & Goldstein, H. (2009). The limitations of using school league tables to inform school choice. *Journal of the Royal Statistical Society: Series A, 172,* 835–851.

Lizzio, A., Wilson, K., & Simons, R. (2002). University students' perceptions of the learning environment and academic outcomes: Implications for theory and practice. *Studies in Higher Education, 27,* 27–52. doi:10.1080/03075070120099359

Marsh, H. W. (1980). The influence of student, course and instructor characteristics on evaluations of university teaching. *American Educational Research Journal, 17,* 219–237.

Marsh, H. W. (1982). SEEQ: A reliable, valid, and useful instrument for collecting students' evaluations of university teaching. *British Journal of Educational Psychology, 52,* 77–95.

Marsh, H. W. (1984). Students' evaluations of university teaching: Dimensionality, reliability, validity, potential biases, and utility. *Journal of Educational Psychology, 76,* 707–754. doi:10.1037/0022-0663.76.5.707

Marsh, H. W. (1987). Students' evaluations of university teaching: Research findings, methodological issues, and directions for future research [Monograph]. *International Journal of Educational Research, 11*(3). doi:10.1016/0883-0355(87)90001-2

Marsh, H. W. (1991). Multidimensional students' evaluations of teaching effectiveness: A test of alternative higher-order structures. *Journal of Educational Psychology, 83,* 285–296. doi:10.1037/0022-0663.83.2.285

Marsh, H. W. (2007a). Do university teachers become more effective with experience? A multilevel growth model of students' evaluations of teaching over 13 years. *Journal of Educational Psychology, 99,* 775–790. doi:10.1037/0022-0663.99.4.775

Marsh, H. W. (2007b). Students' evaluations of university teaching: Dimensionality, reliability, validity, potential biases and usefulness. In R. P. Perry & J. C. Smart (Eds.), *The scholarship of teaching and learning in higher education: An evidence-based perspective* (pp. 319–384). New York, NY: Springer.

Marsh, H. W., & Bailey, M. (1993). Multidimensional students' evaluations of teaching effectiveness: A profile analysis. *Journal of Higher Education, 64,* 1–18. doi:10.2307/2959975

Marsh, H. W., & Cheng, J. H. S. (2008). NSS: Dimensionality, multilevel structure, and differentiation at the level of university and discipline. Retrieved from http://www.heacademy.ac.uk/assets/York/documents/ourwork/research/surveys/nss/NSS_herb_marsh-28.08.08.pdf

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, NY: Agathon.

Marsh, H. W., & Hau, K.-T. (2007). Applications of latent-variable models in educational psychology: The need for methodological–substantive synergies. *Contemporary Educational Psychology, 32,* 151–170. doi:10.1016/j.cedpsych.2006.10.008

Marsh, H. W., Lüdtke, O., Muthén, B., Asparouhov, T., Morin, A. J. S., Trautwein, U., & Nagengast, B. (2010). A new look at the Big Five factor structure through exploratory structural equation modeling. *Psychological Assessment, 22,* 471–491. doi:10.1037/a0019227

Marsh, H. W., Muthén, B., Asparouhov, T., Lüdtke, O., Robitzsch, A., Morin, A. J. S., & Trautwein, U. (2009). Exploratory structural equation modeling, integrating CFA and EFA: Application to students' evaluations of university teaching. *Structural Equation Modeling, 16,* 439–476. doi:10.1080/10705510903008220

Marsh, H. W., & Roche, L. A. (1997). Making students' evaluations of teaching effectiveness effective. *American Psychologist, 52,* 1187–1197. doi:10.1037/0003-066X.52.11.1187

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. doi:10.1037/0022-0663.92.1.202

Marsh, H. W., Rowe, K. J., & 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,* 313–348. doi:10.1353/jhe.2002.0028

Marsh, H. W., Seaton, M., Trautwein, U., Lüdtke, O., Hau, K. T., O'Mara, A. J., & Craven, R. G. (2008). The big-fish–little-pond-effect stands up to critical scrutiny: Implications for theory, methodology, and future research. *Educational Psychology Review, 20,* 319–350. doi:10.1007/s10648-008-9075-6

McKeachie, W. J. (1979). Student ratings of faculty: A reprise. *Academe, 65,* 384–397. doi:10.2307/40248725

McKeachie, W. J. (1997). Student ratings: The validity of use. *American Psychologist, 52,* 1218–1225. doi:10.1037/0003-066X.52.11.1218

McKinnon, K. R., Walker, S. H., & Davis, D. (2000). *Benchmarking: A manual for Australian universities.* Canberra, Australian Capital Territory, Australia: Australian Department of Education, Training and Youth Affairs. Retrieved from http://www.dest.gov.au/archive/highered/otherpub/bench.pdf

Muthén, L. K., & Muthén, B. (2010). *Mplus user's guide.* Los Angeles, CA: Muthén & Muthén.

Parsons, P. G. (1988). The Lancaster approaches to studying inventory and course perceptions questionnaire: A replicated study at the Cape Technikon. *South African Journal of Higher Education, 2,* 103–111.

Patrick, K. (2003, March). The CEQ in practice: Using the CEQ for improvement. In Graduate Careers Council of Australia symposium *Graduates: Outcomes, Quality and the Future,* Canberra, Australian Capital Territory, Australia.

Ramsden, P. (1976). *Course evaluation in higher education* (Unpublished masters thesis). Council for National Academic Awards, Lancaster, United Kingdom.

Ramsden, P. (1979). Student learning and perceptions of the academic environment. *Higher Education, 8,* 411–427. doi:10.1007/BF01680529

Ramsden, P. (1991a). A performance indicator of teaching quality in higher education: The Course Experience Questionnaire. *Studies in Higher Education, 16, 2,* 129–150. doi:10.1080/03075079112331382944

Ramsden, P. (1991b). Report on the Course Experience Questionnaire trial. In R. Linke (Ed.), *Performance indicators on higher education* (Vol. 2, pp. 1–85). Canberra, Australian Capital Territory, Australia: Australian Government Publishing Service.

Ramsden, P. (1992). *Learning to teach in higher education.* London, United Kingdom: Routledge. doi:10.4324/9780203413937

Ramsden, P., & Entwistle, N. J. (1981). Effects of academic departments on students' approaches to studying. *British Journal of Educational Psychology, 51,* 368–383.

Rasbash, J., Steele, F., Browne, W., & Prosser, B. (2005). *A user's guide to MLwiN Version 2.0.* Bristol, United Kingdom: University of Bristol.

Raudenbush, S. W., & Bryk, A. S. (2002). *Hierarchical linear models* (2nd ed.). Thousand Oaks, CA: Sage.

Richardson, J. T. E. (2005). Instruments for obtaining student feedback: A review of the literature. *Assessment & Evaluation in Higher Education, 30,* 387–415. doi:10.1080/02602930500099193

Watkins, D. (2001). Correlates of approaches to learning: A cross-cultural meta-analysis. In R. Sternberg & L.-F. Zhang (Eds.), *Perspectives on thinking, learning, and cognitive styles* (pp. 139–166). Mahwah, NJ: Erlbaum.

Wilson, K., Lizzio, A., & Ramsden, P. (1997). The development, validation and application of the Course Experience Questionnaire. *Studies in Higher Education, 22, 1,* 33–53. doi:10.1080/03075079712331381121

Received February 9, 2010
Revision received February 4, 2011
Accepted March 4, 2011 ■