British Educational Research Journal
Vol. 34, No. 1, February 2008, pp. 91–115

I liked your course because you taught me well: the influence of grades, workload, expectations and goals on students' evaluations of teaching

Richard Remedios*[a] and David A. Lieberman[b]
[a]*Durham University, UK;* [b]*University of Stirling, UK*

(Submitted 2 October 2005; conditionally accepted 14 February 2006; accepted 5 April 2006)

There has been considerable debate as to whether course evaluations are valid measures of teaching quality, or whether students instead reward tutors who give them high grades and assign low levels of work. To assess the factors that influence course evaluations, we measured university students' achievement goals and expectations at the beginning of the semester and also obtained information on grades and workload. Although grades and course difficulty did have a small influence on end-of-semester course ratings, structural modelling revealed that ratings were largely determined by how much students enjoyed or felt stimulated by the course content, which in turn depended on the perceived quality of teaching. Students with a mastery goal were more likely to look forward to the course, and this also contributed to positive course evaluations, but the effect was small. Overall, the results suggested that by far the largest determinant of student evaluation of courses is the quality of the teaching.

Course evaluations remain the primary method used in higher education to gauge how effectively courses are taught. However, the validity of these ratings has been a matter of considerable, sometimes heated, debate (see D'Apollonia & Abrami, 1997; Greenwald & Gilmore, 1997a; Marsh & Roche, 1997; McKeachie, 1997). On the one hand, a substantial body of research suggests that ratings are not a valid measure of teaching quality, because they are biased by factors such as grading leniency (e.g. Powell, 1977; Vasta & Sarmiento, 1979; Worthington & Wong, 1979; Blunt, 1991; Chako, 1983; Greenwald & Gilmore, 1997a, b; Olivares, 2001; Griffin, 2004; also see Feldman, 1976 and Stumpf & Freedman, 1979 for reviews), workload (e.g. Greenwald & Gilmore, 1997a, b; Griffin, 2004) and pre-course expectations (e.g.

*Corresponding author. School of Education, Durham University, Leazes Road, Durham, DH1 1TA, UK. Email: richard.remedios@durham.ac.uk

ISSN 0141-1926 (print)/ISSN 1469-3518 (online)/08/010091-25

DOI: 10.1080/01411920701492043

Remedios *et al.*, 2000; Griffin, 2004). On the other hand, other evidence suggests that these biasing factors actually have little or no effect and that course evaluations do provide a valuable index of teaching quality (e.g. Marsh & Roche, 2000). The current study investigates how factors such as students' pre-course expectations, achievement goals, grades, workload, and perceptions of course difficulty affect how they rate their courses.

The role of grades and workload in students' evaluations of their courses

The positive correlation between course ratings and grades has been interpreted as evidence that students reward instructors who award them high grades. Similarly, the negative correlation between course ratings and workload has been seen as evidence that students prefer courses where they do not have to work hard (Greenwald & Gilmore, 1997a, b; Olivares, 2001; Griffin, 2004). Marsh and Roche (2000), however, have argued that these interpretations are incorrect, and that course evaluations are valid indicators of teaching quality. Good teaching, they suggest, leads to students learning more and therefore attaining higher grades. In this view, the correlation between grades and ratings is actually a tribute to the validity of these ratings as a measure of teaching quality, rather than evidence that student ratings are biased. Marsh and Roche (2000) criticised Greenwald and Gilmore's (1997a) use of structural models to show a grading leniency effect because the models failed to control for student learning. When Marsh and Roche reanalysed Greenwald and Gilmore's data, they found that student learning accounted for much of the variance in the grade–course rating relationship; students who learned more were also more likely to rate courses positively. As for the negative relationship between workload and course ratings observed by Greenwald and Gilmore, Marsh and Roche contended that courses with high workload should lead to students learning more. If amount of learning was controlled, they argued, workload on its own would have only a small effect on course ratings, and re-analysis of the data confirmed this prediction.

To summarise, current evidence suggests that although grades and workload predict how students will evaluate their courses, this might not be because students are rewarding instructors for awarding high grades or for assigning low amounts of work. Quite the contrary, ratings might be an accurate reflection of how well a course is taught. Good teaching, in this view, leads to better learning, and this in turn leads to both good grades and high course ratings.

The role of students' expectations

Research across many areas of psychology has suggested that the way in which people react to events is often strongly influenced by their expectations (e.g. Feather, 1961, 1963a, b, 1966; Berger *et al.*, 1977, 1985). Given the importance of expectations in other situations, it is plausible to assume that how students react to their courses might also be influenced by their expectations prior to beginning the

course. If, for example, students entered a course expecting high grades, they would be more likely to be disappointed if they received average grades. On this assumption, several investigations of course evaluations have measured students' expectations about what grades they would receive (Greenwald & Gilmore, 1997a; Olivares, 2001; Griffin, 2004). However, these expectations have usually been measured at the end of courses, at the same time as students completed their course evaluations.

As pointed out by Remedios *et al.* (2000), grade expectations at the end of a course will not necessarily be the same as those at the beginning of the course. By the end of a course, most students will have already received feedback on several pieces of work, and this feedback may have led to modifications in their expectations. Put another way, grade expectations at the end of a course are probably best seen as students' current predictions of their grades, rather than as a measure of the aspirations with which they began the course. In so far as students' expectations influence how they react to their courses, it is more likely to be their expectations at the beginning of the course that are important, rather than the grades they expect at the end.

To measure students' expectations more accurately, Remedios *et al.* (2000) asked students to complete a questionnaire asking what grade they expected at the beginning of the semester. The second questionnaire, administered at the beginning of the following semester, then asked students to rate how interesting they had found the course and how much they had enjoyed it. The results confirmed the importance of students' expectations in how they reacted to grades. That is, the best predictor of interest and enjoyment was not grades per se but rather the difference between the grades students had expected at the beginning of the course and the grades they actually received—the more students' grades exceeded their expectations, the more they reported enjoying the course.

One limitation of Remedios *et al.*'s study was that it used only two measures of students' reactions to their courses, interest and enjoyment. Thus, while the degree of missed expectations predicted these variables, it was not possible to assess the effect on students' overall evaluations of their courses. One purpose of the current study, therefore, was to provide a more wide-ranging assessment of how students had perceived their courses by including questions about issues such as course organisation and grading.

A second purpose of this study was to explore other aspects of students' expectations about their courses, and whether these other aspects also influenced their reactions to their courses. Marsh (1983, 1987) reported that students' prior interest in a course's topic had a substantial effect on the relationship between grades and course ratings, accounting for one-third of the variance. However, interest was measured during the course and was therefore a post-hoc measure of prior interest and could have been affected by how students were doing at that time. The most appropriate way to ensure that ratings of initial interest are not confounded by experiences during the course (hindsight bias—see, for example, Fischhoff & Beyth, 1975; Fischhoff, 1975) is to assess interest before the course starts. One

questionnaire in our study was administered before students began their courses and this questionnaire included questions about how interesting and enjoyable they expected their courses to be.

The role of students' goals and intrinsic motivation

A further purpose of this study was to examine how students' motives for studying influence their reactions to their courses. A variety of theories have been proposed to describe students' motivations for studying, but we will focus here on two, achievement goal theory and intrinsic motivation theory. Achievement goal theorists have proposed that individuals differ in the goals they adopt in situations involving achievement. In the case of studying, it was initially claimed that these goals fell into two categories, the goal of understanding or mastering a subject, and the goal of performing well and thus looking good in front of others (e.g. Elliott & Dweck, 1988; Dweck & Leggett, 1988; see Pintrich, 2003, for review). Elliott and colleagues subsequently identified a third approach which they labelled performance avoidance, defined as the goal of avoiding performing badly (e.g. see Elliot & Harackiewicz, 1994, 1996; Elliot & Church, 1997).[1]

Intrinsic motivation theorists (e.g. Deci, 1975; Deci & Ryan, 1985) suggest that individuals engage with tasks for either internal reasons (i.e. because they want to) or external reasons (i.e. because of external pressures). One operationalisation of intrinsic motivation is self-reported interest and enjoyment: Students who are intrinsically motivated to study are assumed to be more likely to be interested in their courses and to enjoy them (see Deci, 1975).

Although the two theories differ in detail, they have converged over time and both now emphasise a distinction between students whose goal is to understand their subjects (a mastery goal or intrinsic motivation) and students whose goal is to obtain high grades (a performance goal or extrinsic motivation). The two goals are not incompatible—a student could be interested in both grades and understanding—but research has suggested that differences in the goals students adopt can have important implications for how they react to their courses. Elliot and Church (1997), for example, examined how the achievement goals of 204 undergraduates, measured two weeks into their course, related to their post-course grades and subsequent intrinsic motivation (e.g. fun, interest, enjoyment). Using path analysis, they found that a mastery goal was positively related to intrinsic motivation and a performance-approach orientation was positively related to final grades, whilst a performance-avoidance orientation was negatively related to both intrinsic motivation and final grades (see also Harackiewicz *et al.*, 2000, 2002).

This research suggests that students' goals can strongly influence their grades as well as how much they enjoy their courses, but many questions remain. For example, if students are intrinsically motivated, does this mean that they are more likely to enjoy all their courses, regardless of how well these courses are taught? Conversely, if students are performance oriented, does this mean that they are more likely to evaluate their courses on the basis of the grades they receive, rather than

how well they are taught? By examining the relationship between students' goals and their course evaluations, we hoped to shed some light on these questions.

The role of models

As outlined earlier, one of the goals of this research was to explore the clash between two fundamentally different views of how students evaluate their courses: the validity view, which sees students' evaluations as fundamentally valid measures of how well courses are taught, and the bias view, which sees evaluations as too heavily influenced by factors such as the grades awarded to provide a useful index of teaching quality. Evaluating these perspectives, however, is not straightforward. In the case of grades, we have seen that it is not enough to show that grades influence student evaluations; to properly assess this relationship we need to take into account the context. For example, we need to know whether grades are themselves the product of how much students learn, so that high grades are actually better interpreted as evidence of good learning—and thus of the validity of evaluations—rather than as evidence of bias. In other words, we need to understand the relationship between the many variables that influence evaluations before we can assess whether evaluations provide a valid measure of teaching quality.

Where many independent variables influence one or more dependent variables, it can be very difficult to disentangle the relationships, and doing so requires the construction of theoretical models to chart the linkages. The construction of such models has been made considerably easier by the emergence of statistical modelling tools such as LISREL and AMOS. However, the adage of 'garbage in, garbage out' still captures an important truth about model construction, as seemingly subtle differences in how models are specified in these analyses can lead to dramatically different conclusions. For example, using structural modelling, Greenwald and Gilmore (1997a, b) concluded that workload and grading leniency both had substantial effects on students' evaluations of their courses. However, Marsh and Roche (2000) criticised Greenwald and Gilmore's models on the grounds that they had not controlled for the amount students learned. When Marsh and Roche re-analysed the data using a model that controlled for student learning, they found that grading leniency had no effect, and workload had the opposite effect to that found by Greenwald and Gilmore. It is not so much model testing that is the problem, it is model specification.

In the present study, we measured students' goals and expectations at the beginning of each course, and then their evaluations of the course at the end. We also obtained information on students' grades and their rating of each course's workload. To help us in developing an appropriate model for analysing all this data, we engaged in the following process. In the first stage, we focused on the achievement goal questionnaire; this had 13 questions, and we examined whether students' answers seemed to tap some smaller number of intervening variables or latent constructs. Similarly, we analysed the underlying structure of our course evaluation questionnaires and our questions about expectations. Finally, we

developed and tested a structural model that explained how all these variables interrelated.

In assessing the role of grades, we used students' actual grades, rather than the mid-course expected and relative grades used by Greenwald and Gilmore (1997a, b). For workload, we used students' estimates of the number of hours they studied and also their rating of the difficulty of each course.

Method

Participants

Some 765 students studying Psychology at a Scottish university were asked to complete a questionnaire as they waited to register in September 2001. Of these 765 students, 722 completed a questionnaire. Of these 722 students, 479 completed a second questionnaire when they registered at the beginning of the following semester (February 2002). First-year students took an Introductory Psychology course, second-year students took a Social Psychology course whilst third-year students studied between one and three courses, drawn from Cognition, Perception, Learning and Psychopathology. From a selection of 12 different units, final-year students took two half-credit electives.

Because some students studied more than one Psychology course, the data for this study is based on the number of valid sets of questionnaires students completed, rather than on the number of students who participated. A valid set was defined as the completion by a student of both questionnaires for one course. First-year students completed 206 valid sets, second-year students 107, third-year students 213 and fourth-year students 84. In total, 610 valid sets were used in the final analysis.

Materials

Pre-course questionnaire. An introductory page gave background details of the study; this emphasised the voluntary nature of participation and asked participants to record their student identification number on the questionnaire if they were willing to take part. The second page contained general questions about the student and his or her goals, and the third page contained questions relating specifically to the courses they would be taking.

The first section of the general questions consisted of 18 questions drawn from a goal-orientation questionnaire devised by Elliot and Church (1997). Participants were asked to rate their agreement, on a 7-point Likert scale, with statements such as:

- It is important to me to be better than other students.
- It is important for me to understand the content of my courses as thoroughly as possible.

Possible responses were numbered 1 to 7; response 1 was labelled 'Not true of me' and response 7 was labelled 'Very true of me.' These questions were followed by the

following supplementary question, which asked students to weigh their desire for good grades against their desire for understanding:

- For university courses, one possible goal is to understand the material, another is to get a good grade. What is the relative importance of each to you? Specifically, where on the following scale would you locate your feeling, between being primarily concerned with grades at one extreme (the right) and primarily concerned with understanding at the other (the left).

These were again on a 7-point Likert scale; the left-most choice was labelled 'Primarily understanding,' the middle choice was labelled 'Equal importance,' and the right-most choice was labelled 'Primarily grades.' This was followed by two questions asking students to record their age and gender.

In the same questionnaire, students were asked about their expectations concerning the psychology courses they would be taking that semester. For each course they were asked the following questions:

- How much are you looking forward to taking this course?
- How interested are you in studying Psychology?
- How interesting do you expect this course to be?
- How enjoyable do you expect this course to be?
- How much difficulty do you think you will have understanding the material in this course?

Each question was followed by the 7-point Likert scale with 1 labelled 'Not very much' and 7 labelled 'A lot.' Students were then asked how much work they expected to do for the course. This question read, 'On average, how many study hours per week do you think you will need to spend on this course? (study hours include time spent preparing for class meetings, tutorials, exams and preparing/writing essays)' and the answer options were, under 2, 2–3, 4–5, 6–7, 8–10, 11–13, 14–16, 17–19, 20–22, 23–25, 25+.

Finally, students were asked to indicate what grade they expected to get for each course. Students in years 2, 3 and 4 were asked to select a grade on the University's grading scale. This contains 5 categories (1, 2.1, 2.2, 3 and 4), with each of the first four categories further divided into three subcategories. Students in the first year, who were not yet familiar with the university system, were presented with the options A+, A, A−, B+, B, B−, etc., a grading system used in secondary education in the UK and more likely to be familiar to first-year students.

Course evaluation questionnaire. Students were asked to fill in a questionnaire for each of the psychology courses they had taken the previous semester. Each questionnaire began by asking students to fill in their registration numbers, and also to indicate which of their courses they were evaluating.

Our choice of which questions to use in the remainder of the form was influenced by several considerations. One was that students were going to be asked to complete several questionnaires at different times throughout the semester. In our previous (longitudinal) studies (e.g. Remedios *et al.*, 2000), we had noticed that students had

become increasingly reluctant to complete what seemed to them to be 'yet another questionnaire'. Our study was entirely voluntary and relied on students' goodwill to complete the questionnaires. Also, traditionally, course evaluations are distributed towards the end of a course or at the final examination, and time is therefore specifically allocated for students to complete the questionnaires and students are encouraged to do so. In our study, students were approached in a variety of ways (e.g. whilst they waited in line for registration, prior to a class, via a letter to their home address). We therefore wanted a questionnaire that would be user-friendly, but without undermining its psychometric properties. To achieve this, we designed our questionnaire so that all the questions could be fitted on to one side of an A4 sheet of paper.

We based our questionnaire on Stringer and Irwing's (1998) Teaching Effectiveness Survey (TES), because this instrument was a composite of several other questionnaires and had been used for a UK population. The TES contained 25 items that fell into five main categories: teaching quality, feedback and support, learning, workload and overall evaluation. We selected 11 items for inclusion in our questionnaire and added a further question about teacher enthusiasm. The decision to add a question on enthusiasm was based on a finding from Remedios *et al.* (2000). As part of their study, they used a focus group ($n=6$) selected from students who had completed the questionnaires. When these students were asked what factors influenced their enjoyment of courses, the most frequent response was how enthusiastically the course was taught (see also Perry *et al.*, 1979; Marsh, 1987; Griffin, 2004).

Each question was followed by the 7-point Likert scale where 1 represented the most negative rating and 7 the most positive. The questions used, and the headings under which they were presented, were as follows:

Teaching Quality:

- On the whole, how well do you think this course was *organised*? (1=not well, 7=very well)
- On the whole, how well do you think this course was *taught*? (1=not well, 7=well)
- On the whole, how *enthusiastically* was this course delivered by the lecturer(s) (1=not enthusiastically, 7=very enthusiastically)

Feedback and Support:

- How would you rate the quality of the feedback you received for your assignments (e.g. lab reports, essays)? (−3=not useful, +3=very useful)
- On the whole, how *fairly* do you think your assignments have been graded? (1=not fair, 7=very fair)
- How would you rate the level of *support* you received on this course (e.g. information on course website; help from lecturers/tutors, etc.) (1=very poor, 7=very good).

Personal Experiences of Course Content:

- How *intellectually stimulating* did you find the course? (1=not stimulating; 7=very simulating)

- How *interesting* did you find the course?(−3=not very much, +3=a lot)
- How *useful* do you think the material learned in this course will be to you in terms of your overall degree? (−3=not useful, +3=very useful)

Course Difficulty:

- On the whole, how *difficult* did you find this course? (−3=very difficult, +3=very easy)

Overall Evaluation:

- How much did you *enjoy* the course? (−3=not very much, +3=a lot)
- Would you recommend this course to another student? (−3=definitely not, +3=definitely yes)

Finally, students were asked two questions concerning their workload and grade:

- On average, how many study hours per week did you spend on this course? (study hours=time spent preparing for lectures, tutorials, labs, exams and preparing/writing essays)
- In terms of the quality of your own work, what final grade do you think you *deserved*?

The workload question was followed by 11 options, again ranging from under 2 to over 25; the grade question was followed by options ranging from 4 (fail) to 1A (high first class degree).

Procedure

The study took place in the period from September 2001 to February 2002 and involved all students studying Psychology during the autumn semester. Students were administered two questionnaires, one at registration in the autumn and one at the beginning of the following semester. On both occasions, students were asked if they would be willing to participate in the study as they waited in a queue to register. They were asked to deposit their questionnaires, once they had completed them, in one of several boxes marked 'completed questionnaires.' In addition, a helper stood at the exit of the registration room and as the students left the room, they were politely asked if they had completed a questionnaire. Four versions of the pre-course questionnaires were used, depending on which year the student was in; these versions were colour-coded. All students received the same questionnaire at the beginning of the following semester; they were asked to complete a separate questionnaire for each psychology course they had completed the previous semester.

Because students had indicated their registration numbers on the questionnaires, we were able to identify which students had not completed questionnaires. For both stages of the study, students who did not complete a questionnaire were sent a letter asking if they would be willing to complete a questionnaire.

Results

The analysis of the data proceeded in three stages: exploratory factor analysis (EFA) and confirmatory factor analysis (CFA) of the achievement-goal questionnaire used in questionnaire one; exploratory and confirmatory factor analysis on the course evaluation questionnaire; and structural modelling to determine the relationships between course evaluation ratings and factors such as achievement motivation, grades, study hours, perceived difficulty, and pre-course expectations.

Stage One: exploratory and confirmatory factor analysis for the achievement-goal questionnaire

Selection of fitness criteria. Before describing the analysis for the achievement-goals questionnaire, it is probably useful for us to spell out the logic behind our decision to use certain goodness-of-fit statistics.

All our confirmatory factor analysis was conducted using the statistical package AMOS (version 4). This package allows the testing of many possible models; to determine which model or models provides the best fit of the data, a variety of goodness-of-fit statistics can be used. Marsh and Roche (2000) used the Relative non-centrality index (RNCI), the Tucker-Lewis index (TLI), and the likelihood ratio test (χ^2) statistic to determine degree of fit. However, AMOS output does not produce the RNCI statistic so we could not include that in our output. For the TLI, Marsh and Roche set the fit limit at .9. However, Hu and Bentler (1999) recommend the limit be set at .95 for large sample sizes, and this was the value we used.

In addition to the indices used by Marsh and Roche (2000), we used several others. The primary additional measure was the Root Mean Square Error of Approximation (RMSEA). Byrne (2001) highlights the increasing acceptance of the RMSEA as the statistic most relevant to model fitness. There is, however, some difference of opinion as to what constitutes a good fit. For example, Byrne (2001) recommends that fits$<$.05 can be classed as good, whilst Hu and Bentler (1999) have suggested values$<$.06 and Browne & Cudek (1993) suggest values as high as$<$.08. We preferred to be conservative and accepted only models where RMSEA was$<$.05. To support the RMSEA, one useful statistic that AMOS produces is labelled PCLOSE. This statistic tests the confidence limits around the RMSEA to ensure these limits are not overly wide. Some models may pass the RMSEA fit statistic but the actual distribution of values may deviate in a way that would lead us to question whether the model is in fact a good fit (see MacCallum *et al.*, 1996). Because PCLOSE helps to limit possible Type II errors, we chose to use it together with RMSEA. Following Joreskog and Sorbom's (1996) recommendation, we accepted models only where the value of PCLOSE exceeded .50.

We also used two other measures. One was the Incremental Fit Index (IFI), which Bollen (1989) recommends values to be$>$.95 to indicate a good fit. We also used Hoelter's Critical N. This statistic focuses on the adequacy of sample size rather than on model fit. Specifically, its purpose is to estimate the sample size that would be

sufficient to yield an adequate model fit for a χ^2 test. Hoelter (1983) suggested that a value in excess of 200 is indicative of a model that adequately represents the sample data. This statistic is particularly useful because models which pass the IFI and TLI fit limits can often fail the Hoelter criteria. In this sense, the Hoelter applies a degree of conservativeness in helping to decide which models should be accepted as good fits.

Our general preference for conservative estimates is based on the fact that in confirmatory factor analysis it is possible for many models to pass fitness tests, and, as we explain later, we tested many models that seemed theoretically plausible. By employing a range of criteria, we hoped to minimise the possibility of a Type II error. We therefore used χ^2, TLI, IFI, RMSEA, PCLOSE and Hoelter's CN in evaluating all of our models. For those models that passed all our fitness criteria, we used one of two methods to assess which model was superior. If the degrees of freedom were the same, we simply compared χ^2 and selected the model with the lower χ^2. If the degrees of freedom were different, we took the difference in χ^2 and the difference in degrees of freedom and used look-up tables to examine whether the difference was significant, e.g. difference in $\chi^2=4$, difference in df=1. C.V. χ^2 (1)=3.84, therefore the difference would be significant and we would choose the model with the lower χ^2.[2]

Exploratory factor analyses. The achievement-goal questionnaire was developed by Elliot and Harackiewicz (1994) and validated in subsequent studies (e.g. Elliot & Harackiewicz, 1996; Elliot & Church, 1997). To confirm whether its properties were appropriate for our sample, we conducted the following analyses.

Firstly, in a similar fashion to Elliot and Church (1997), we submitted the 18 questionnaire items to principal components analysis with a varimax rotation. Unlike Elliot and Church (1997), whose solution produced three components with eigenvlaues>1, our initial solution produced four components, accounting for 61.9% of the total variance. (The eigenvalues and percentage of variance accounted for by each component were as follows: component one $\lambda=4.2$ [23.2%]; component two $\lambda=3.4$ [19.1%]; component three $\lambda=2.4$ [13.5%], component four $\lambda=1.1$ [6.1%].) This initial solution appears in Table 1.

Table 1 shows that for the initial four-factor solution, 14 items fell into the same categories as defined by Elliot and Church (1997). Items Performance-Approach (P) 15, Mastery (M) 14, M17 and Performance-Avoidance (A) 18, however, did not fall cleanly into the predicted factors. We forced a three-factor solution to help confirm to which construct these four items most appropriately belonged. The subsequent three-factor solution appears on the right-hand side of Table 1 and provides some clarification of the structure of the constructs. The difference between the primary and secondary loadings for items M14 and M17 were now .43 and .39 respectively. This suggested that it was appropriate to place M14 and M17 in the Mastery factor. The appropriate location of two items, however, remained problematic: A18, because its primary loading was low, and P15, because it loaded almost equally on two factors.

Confirmatory factor analyses. We began by testing two models, one that contained the full 18-item achievement-orientation questionnaire (model A1, χ^2 (132)=512.7, TLI=.977, IFI=.983, RMSEA=.07, PCLOSE=.000, HCN=190) and one the reduced 16-item questionnaire with questions A18 and P15 omitted (model A2, χ^2

Table 1. Three and four (rotated) component solutions for the 18-item achievement-orientation questionnaire. Primary factor loadings are in bold and secondary loadings are in italics if they were greater than .10.

4-component				Questionnaire items (P=Performance, M=Mastery, A=Avoidance. Numbers represent the number of the question on the questionnaire)	3-component		
1	2	3	4		1	2	3
.83			*.11*	P1. It is important to me to be better than other students.	.83		
.90				P4. My goal in my courses is to get a better grade than most of the other students.	.90		
.78				P7. I am striving to demonstrate my ability relative to others in my courses.	.77		*.10*
.85				P10. I am motivated by the thought of outperforming my peers.	.85		
.85				P13. It is important to me to do well compared to others in my courses.	.84		*.11*
.42		**.46**		P15. I want to do well to show my ability to my family, friends and others.	*.40*		**.47**
	.82	*.14*		M2. I want to learn as much as possible from all my courses.		**.80**	*.18*
	.81			M5. It is important for me to understand the content of my courses as thoroughly as possible.		**.80**	*.12*
	.83		*.11*	M8. I hope to have gained a broader and deeper knowledge when I have completed my courses.		**.82**	*.15*
.14	.72			M11. I desire to completely master the material presented in my courses.	*.11*	**.71**	
	.45		**.63**	M14. I prefer material that arouses my curiosity, even if it is difficult to learn.	*.17*	**.60**	
	.47		**.70**	M17. I prefer a course that really challenges me so I can learn new things.	*.14*	**.63**	
	.17	**.82**		A3. I often think to myself, 'What if I do badly in my courses?'		*.14*	**.82**
	.15	**.84**		A6. I worry about the possibility of getting a bad grade my courses.		*.10*	**.84**
.25		**.64**		A9. My fear of performing poorly is often what motivates me.	*.24*		**.65**
		.55	*.29*	A12. I just want to avoid doing poorly in my courses.			**.54**
		.59		A16. I'm afraid that if I ask a 'stupid' question, the lecturer might think I'm not very clever.			**.59**
		.24	**.50**	A18. I wish my courses were not graded.		*(−.31)*	**.21**

(101)=350.0, TLI=.983, IFI=.987, RMSEA=.06, PCLOSE=.001, HCN=219). Model A2 proved to be a better fit than model A1, and further analysis revealed this difference to be significant (χ^2 (31)=162.7, $p<.001$). However, model A2 did not pass all the fitness criteria, and this suggested that there was a better solution. Inspection of Table 2 shows that three other items showed cross-loadings against other factors. For example, M14 and M17 loaded on the performance construct and A9 also loaded on the performance construct. Initially, we were prepared to accept

Table 2. Two and three factor (rotated) solutions for the 13-item course evaluation questionnaire. Primary pattern matrix factor loadings are in bold and secondary loadings are in italics if they were greater than .10.

3-factor			Course Evaluation Questionnaire (CEQ) items	2-factor	
1	2	3		1	2
	.89		Q1. On the whole, how well do you think the course was organised?		**.90**
.11	**.85**		Q2. On the whole, how well do you think the course was taught?		**.85**
.11	**.72**		Q3. On the whole, how enthusiastically was the course delivered by the lecturers?		**.72**
	.56		Q4. How would you rate the quality of the feedback you received for your assignments (e.g. lab reports, essays)		**.56**
.14	**.34**	*.14*	Q5. On the whole, how fairly do you think your work was graded?	.20	**.34**
	.60		Q6. How would you rate the level of support you received on the course (e.g. course handouts, information on course website, help from lecturers/tutors, etc.)		**.60**
.84		*−.14*	Q7. How intellectually stimulating did you find the course?	**.80**	
1.05		*−.14*	Q8. How interesting did you find the course?	**.95**	
.72			Q9. How useful do you think the material learned in this course will be to you in terms of your overall degree?	**.77**	
	.12	*.14*	Q10. On the whole, how difficult did you find this course?	NA	NA
.78		*.31*	Q11. How much did you enjoy the course?	***.92***	
.69		*.37*	Q12. Would you recommend this course to another student?	***.86***	

that cross-loadings in excess of .4 represented discrimination between constructs. Given the failure of model A2 to meet our fitness criteria, we increased the criterion for discrimination to .6. This produced model A3 in which questions A9, M14 and M17 were excluded. This model performed well (χ^2 (62)=134.8, TLI=.995, IFI=.995, RMSEA=.04, PCLOSE=.832, HCN=368). Further analysis revealed that model A3 was not only a significantly better fit than model A2 (χ^2 (50)=234.4, $p<.001$) but was also the only model to pass all our fit criteria.

As a final test of the appropriateness of model A3, we ran reliability analyses on the three factors. For the 5-item performance-approach construct, Cronbach's α was .90; for the 4-item mastery-approach construct, α was.83; and for the 4-item avoidance construct, α was .69. The three constructs therefore proved to be reliable.

To summarise, despite our initial PCA solution revealing four factors instead of the three found by Elliot and Church (1997), subsequent investigation using confirmatory factor analysis led us to accept a reduced version of the achievement-goal constructs developed by Elliot and Church.

An 'anticipation' construct. The pre-course questionnaire also contained four questions concerning pre-course interest, pre-course anticipated enjoyment, how much the student was looking forward to taking the course, and how much interest

the student had in Psychology as a whole. For these four questions, we ran a principal components analysis. This produced one component with an eigenvalue>1 (λ=3.3) accounting for 82.0% of the total variance. Reliability for these for four items was high (=.93). We labelled this construct Anticipation.

Stage two: exploratory and confirmatory factor analysis for the course evaluation questionnaire

Exploratory analysis. As noted earlier, our course evaluation questionnaire was drawn from questionnaires used in previous research, i.e. Greenwald and Gilmore (1997a) and Stringer and Irwing (1998). Based on earlier research by Marsh and Roche (2000), Stringer and Irwing (1988) and Greenwald and Gilmore (1997), our a priori hypothesis was that our course evaluation questionnaire would split into five factors/dimensions. In a similar fashion to Stringer and Irwing (1998), we submitted our 12 items to exploratory factor analysis using maximum likelihood extraction followed by direct oblim rotation (delta=0). A correlational analysis of all the items showed them to be highly intercorrelated. Tabanchik and Fidell (1996) recommend using the pattern matrix to discriminate correlated factors because this matrix omits shared variance (p. 653). The results of the pattern matrix appear in Table 2.

Table 2 reveals that the pattern matrix produced a three-factor solution with eigenvlaues>1 accounting for 68.8% of the total variance. (The eigenvalues and percentage of variance accounted for by each component were as follows: component one λ=6.4 [48.0%], component two λ=1.3 [10.8%], and component three λ=1.0 [8.0%].) Table 2 reveals that the component structure of the three-factor model (on the left-hand side of the table) was somewhat cluttered. For example, CEQ5 had a low primary loading on factor 2 (.34) but also had low but similar secondary loadings (.14) on the two other factors. CEQ10 had an extremely low loading on two factors whilst question CEQ13 did not load on any factor at all.

To address these problems, we trimmed the questions on study hours (CEQ13) and course difficulty (CEQ10) from the analysis and conducted a second factor analysis, again using maximum likelihood estimates with oblim rotation (delta=0). This solution produced two components with eigenvlaues>1 accounting for 64.8% of the total variance. (The eigenvalues and percentage of variance accounted for by each component were as follows: component one λ=6.5 [53.8%], and component two λ=1.3 [10.9%].) The two-factor model (on the right-hand side of the table) revealed a much more stable structure with questions 1, 2, 3, 4, and 6 loading on factor 2 and questions 7, 8, 9, 11, and 12 loading on the second factor. There were no secondary loadings associated with the primary ones. Question 5 remained anomalous and was removed.

Confirmatory analysis. Having established which questions combined to form stable constructs, we next investigated how these constructs related to one another. In the next set of explanations, the symbol → represents the direction of the path and implies causality.

Stringer and Irwing (1998) reported that the dimensions of teaching effectiveness could be broken down into three stages, teaching effectiveness (e.g. teaching quality) → course characteristics (e.g. feedback, course integration, work overload) → stimulation/learning → overall evaluation. However, our exploratory analysis suggested that our questionnaire only broke into two constructs, one which seemed to amalgamate Teaching Quality and Support/Feedback (factor 1 in our two-factor model), and another which seemed to combine Learning and Overall Evaluation (factor 2).

We first tested the fitness of this two-factor model. Despite the relatively clear two-factor structure suggested in the exploratory phase of the analysis (see Table 2), the subsequent model in the confirmatory analysis (labelled Model B1) phase proved to be a poor fit, passing none of our fit criteria (χ^2 (34)=345.8, TLI=.939, IFI=.963, RMSEA=0.498, PCLOSE=.000, HCN=86).

Whilst this outcome is surprising, there is other evidence that potentially robust-looking EFA models do not always translate into satisfactory CFA ones. For example, Elliot and Thrash (2002, p. 808) report a similar finding to ours, as their initial EFA suggested a two-factor solution but their subsequent CFA suggested a four-factor one. Thus, the phenomenon of robust-looking factor breakdowns in EFAs translating into something different in CFA is not unusual. Instead, such findings point to the need to conduct careful CFA, even when the EFA seems highly convincing.

Having established that the two-factor model was probably not appropriate, we then examined the best path structure for the four factors. The first model we examined, which we called B2, was the one suggested by Stringer and Irwing (1998). Their model had a linear structure in which teaching effectiveness influenced course effectiveness, which in turn influenced stimulation/learning, which finally influenced students' overall evaluations. Because we used different questions, their terms do not always seem optimal as summaries of our questions, and so we have chosen somewhat different labels for our four factors. As illustrated in Figure 1, we have suggested that Teaching Quality influences ratings of Support; high levels of Support in turn promote greater student Involvement with course material, and this in turn influences Overall Evaluation. This model (χ^2 (32)=181.9, TLI=.969, IFI=.982, RMSEA=0.088, Pclose=.000, HCN=155) was a significantly better fit than model B1 (χ^2 (2)=163.9 $p<.001$), but it failed our RMSEA, PCLOSE and Hoelter criteria.

The next model we tested (B3, see Figure 2) suggested that the Teaching Quality → Involvement relationship was not mediated by Support; instead, Teaching Quality was linked directly to Involvement as well as Support. According to this model, good teaching influences involvement with course material and ratings of Support separately. In other words, the path from Teaching Quality to Involvement does not necessarily need to be mediated by Support—good teaching can result in student involvement even in the absence of high quality Support. Model B3 fitted the data relatively well (χ^2 (31)=85.8, TLI=.988, IFI=.993, RMSEA=0.054, PCLOSE=.299, HCN=320) but like B2 it failed our RMSEA, PCLOSE and Hoelter criteria.

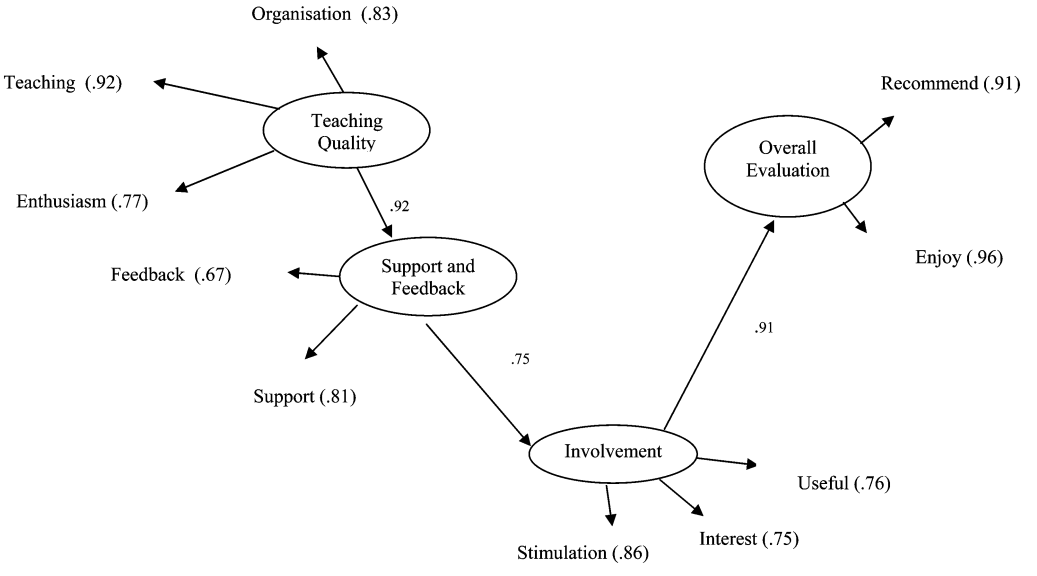

Figure 1. Stage three. Replication of Stringer and Irwing's (1998) four-stage solution for the course evaluation instrument. (Numbers between latent constructs are standardised regression coefficients. Numbers feeding from latent constructs are the observed variables that make up that latent variable together with their standardised regression coefficients)

Byrne (2001) suggests that nearly-fitting models can be honed using post-hoc methods. For example, both Marsh and Roche (2000) and Stringer and Irwing (1998) used LISREL's modification indices to redefine their models, indices that are also available with AMOS. However, some of our data was missing and AMOS does not supply modification indices when data is missing. Rather than unnaturally reducing standard deviations by using techniques to estimate missing data just to review modification indices, we instead completed post-hoc restructuring of the model based on theoretical reasonableness. For example, we looked at whether a model in which Support → Teaching Quality → Involvement→ Overall Evaluation was tenable, but this model proved to be significantly worse. Indeed, of the many theoretically reasonable models we assessed, B3 remained the best fitting one. It was this model, therefore, that we carried into the next stage of the analysis.

Stage three: effect of final grades, study hours and perceived course difficulty on students' evaluations of their courses

The first stages of the analysis determined the dimensional structure of the achievement-goal questionnaire, the dimensional structure of the course evaluation questionnaire and the construct of anticipation, and ensured that the latent constructs in the full structural model reflected the appropriate variables. The final stage concentrated on the main research question, namely, determining how variables that could potentially bias course ratings—final grades, study hours

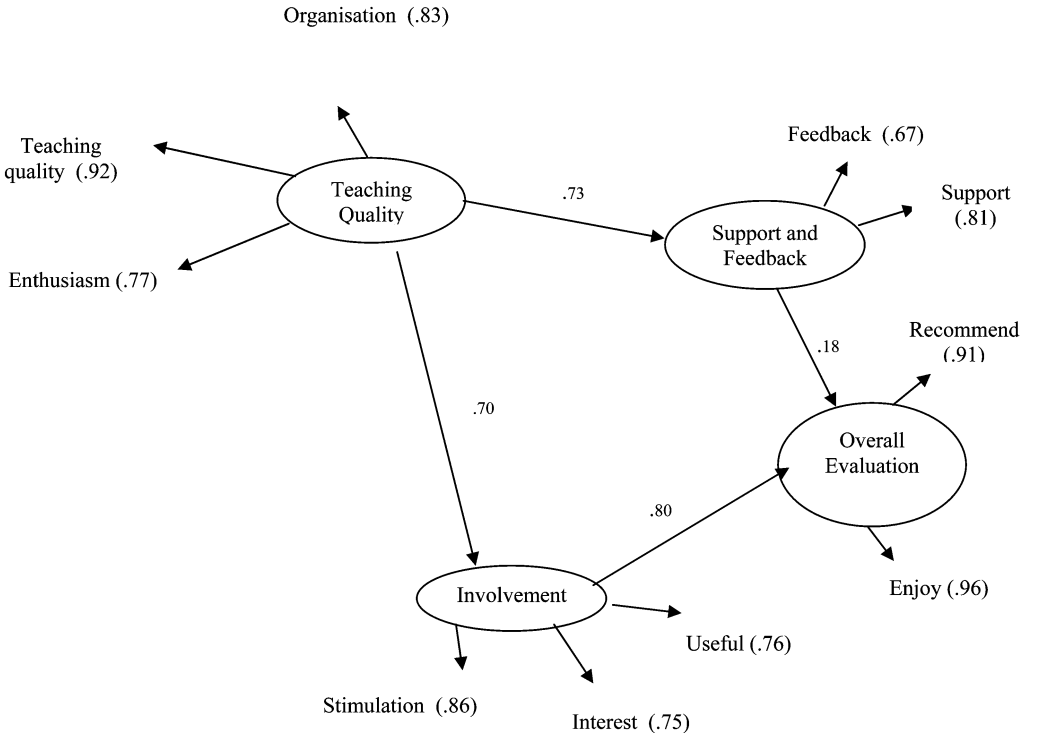

Figure 2. Three-stage solution for the course evaluation instrument where the path from Teaching Quality to Overall Evaluation is mediated by *both* Support and Feedback and Involvement. (Numbers between latent constructs are standardised regression coefficients. The terms not enclosed within the latent variables are the observed variables that make up that latent variable together with their standardised regression coefficients)

(workload), course difficulty, students' goals and expectations—actually influenced students' evaluations.

Our goal was to identify the best model for mapping the three observed variables (final grades, study hours and course difficulty) and two latent variables (anticipation and mastery) against all four dimensions of the course evaluation questionnaire (i.e. teaching quality, involvement, support , and overall rating). There was a very large number of possible models, and our first step in evaluating them was to consider their theoretical plausibility. For example, for one of the models the path from study hours to support was removed a priori because study hours should not predict how individuals rate levels of support. Once a model was accepted as theoretically plausible, we checked if the path coefficients were significant; any paths that were not significant were removed. Finally, we assessed how well the resulting model fitted the data. Models where the fit was poor were excluded, and we then compared the remaining models to see if they differed significantly in fit. The best fitting model (χ^2 (178)=413.4, TLI=.986, IFI=.989, RMSEA=0.047, PCLOSE=.825, HCN=310) appears in Figure 3. In this model, the main determinant of a course's overall evaluation was students' feeling of

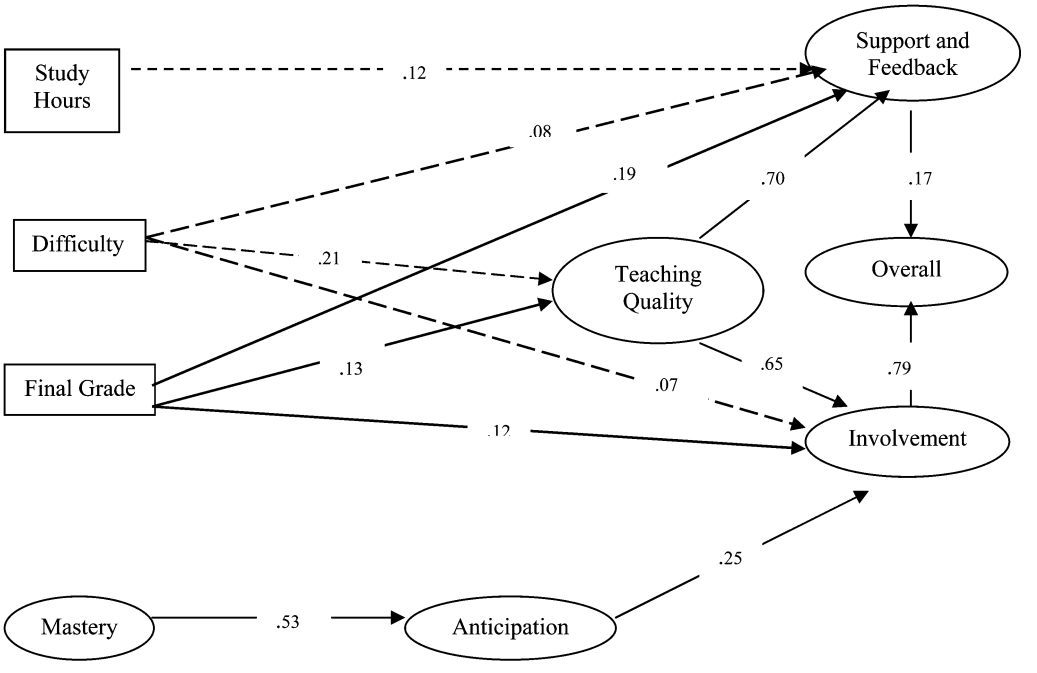

Figure 3. Stage three. The influence of the observed variables of Study Hours, Difficulty and Final Grade, and of the latent variables of Anticipation and Mastery on the four dimensions of the course evaluation questionnaire. (Numbers next to paths are standardised regression coefficients)

involvement with the course material, and this in turn was primarily determined by Teaching Quality. The other main determinant of the overall evaluation was students' feelings of Support, and again this was primarily determined by Teaching Quality.

The variables that have sometimes been suggested to bias student ratings—study hours, perceived difficulty, and final grade—affected overall evaluations only indirectly, via their effect on student feelings of involvement with the course material, and students' perceptions of Teaching Quality and Support. Moreover, the impact of all of these variables was small, and it was also small in all the other models we tested. The least predictive variable was Study Hours, which only predicted ratings of Teaching Quality ($r^2=1\%$). Perceived Difficulty predicted Teaching Quality, Support, and Involvement, but in all cases the effect was small ($r^2=4\%$, 1%, and less than 1%, respectively). Similarly, final grades accounted for only a small proportion of the variance in Involvement ($r^2=1\%$), Teaching Quality ($r^2=2\%$) and Support ($r^2=4\%$).

The other variable that influenced course evaluations in this model was students' desire to master the course material. We tested the influence of the full range of achievement-goals in many models, but only the mastery-goal construct appeared in the fully fit model. Figure 3 shows that having a Mastery orientation influenced how eagerly students anticipated their courses ($r^2=28\%$), and anticipation in turn had a modest effect on how much students felt involved in the course ($r^2=6\%$).

Summary of findings from the EFA, CFA and structural modelling phases of the analysis

Our overall findings can be summarised as follows.

- Exploratory and confirmatory analysis of our achievement-goal questionnaire revealed three underlying factors.
- Analysis of our course questionnaire revealed a four-dimensional structure. This model was similar to that found by Stringer and Irwing (1998), though the relationship between the paths differed somewhat.
- In this model, B3, the main determinants of Overall Evaluation were Support and Involvement, and these in turn were largely determined by Teaching Quality.
- A separate analysis revealed that grades, study hours and perceived difficulty also influenced course ratings, but these influences proved to be small.
- The only achievement goal that significantly influenced course evaluations was Mastery; the influence was indirect and relatively small.

Discussion

This study had two primary purposes, one general and one specific. The general one was to develop a model that that would allow us to understand the factors that influence student evaluations' of teaching, and the complex relationships between these factors. The specific one was to evaluate whether course evaluations provide a valid measure of teaching effectiveness, or, as has sometimes been proposed, whether other factors such as grades and workload influence evaluations so strongly that these evaluations can no longer be viewed as a valid index of teaching quality.

Because conclusions about the impact of variables such as grades depend critically on the model used, we will begin by addressing some of the issues raised by our efforts to construct a model. One crucial step involved analysing the dimensional structure of our course questionnaire. This questionnaire was largely derived from an earlier one developed by Stringer and Irwing (1998), and we, like them, found that students' responses reflected four underlying factors. However, our best-fitting model of how these factors interrelated differed somewhat from that proposed by Stringer and Irwin. They found a linear structure in which Teaching Quality → Support → Involvement → Overall Evaluation (model B2). We, on the other hand, found that Teaching Quality could affect Involvement and Support independently, with both then influencing Overall Evaluation.

One possible reason for this difference was that our questionnaire incorporated only about half of the questions from theirs, and we also added a question on teacher enthusiasm. Whether or not this is the reason for the different outcomes, the optimal path structure that emerged in our model seems to us more plausible. In their model, high quality teaching could not lead to students finding the course content interesting and stimulating (Involvement) unless this teaching also involved high levels of feedback and support. In our model a teacher might be able to inspire students even if this teacher was not good at providing feedback, and this seems

more realistic. This interpretation supports a multidimensional interpretation of course evaluations, because it suggests that students will not necessarily rate courses highly across all dimensions. Even if a course is well taught, our findings suggest that students are able to recognise and rate Support independently of their interest in the course material.

Having established the dimensional nature of the course evaluation questionnaire, the next stage of our analysis investigated how much overall ratings were influenced by study hours, course difficulty, grades, and achievement goals. We should note that we tested every model that seemed theoretically reasonable, and where it appears that paths are missing, this is because the path coefficients were non-significant and/or because there was a better-fitting overall model. In other words, Figure 3 represents the best fit of the data, and it is this model we comment on.

The first question is whether the model supports a validity or a bias hypothesis. The target latent variable Overall Evaluation was derived from two questions, 'On the whole, how much did you enjoy this course' and 'Would you recommend this course to another student?' Figure 3 reveals that none of the variables considered to be biases—study hours, perceived difficulty or final grade—directly predicted this overall evaluation, and even their indirect effects (via Teaching Quality and Involvement) were small. Grades, for example, accounted for only 2% of the variance in Teaching Quality and 1% of the variance in Involvement. Moreover, even these small effects in some cases supported an interpretation in terms of validity rather than bias. In particular, the fact that Difficulty and Study hours influenced perceptions of Teaching Quality actually favours a validity interpretation because, critically, the path coefficients were positive. According to a bias interpretation, students who felt a course was difficult should have rated it more negatively; in fact, they rated it more positively. This finding needs to be interpreted cautiously because of the relatively small number of courses in our survey: Perhaps the more difficult courses in our survey just happened to be the most stimulating. Whatever the explanation, our data provide no support for the view that students penalise instructors whose courses are demanding; in so far as there was an effect, it involved students rating difficult courses more positively.

Overall, our data suggest that students' ratings of courses are largely determined by the degree to which they feel involved, as measured by the extent to which they find their courses stimulating, interesting and useful. In turn, this sense of involvement largely depended on how well students thought a course was organised and taught. Factors such as grades and course difficulty seemed to play at most a very small role.

The small effect of grades in this study parallels other findings in the literature. Marsh and Roche (1997) reviewed a number of studies of course evaluations and suggested that the best estimate of the proportion of variance accounted for by grades is around 4%. If anything, the effect in our study was even smaller, with grades accounting for 2% of the variance in Teaching Quality and 1% of the variance in Involvement. Anecdotal evidence suggests that many instructors believe that high course ratings are often achieved unfairly by overly generous marking. It is

impossible to rule out the possibility that this does occur in some cases, but there is now substantial converging evidence that grades normally have only a small impact on how students evaluate their courses. Students may be fairer, and more perceptive, than we sometimes realise.

Marsh and Roche (2000) strongly advocate taking into account the multi-dimensionality of course evaluations and Figure 3 also supports this position. For example, if we had used only our measure of Overall Evaluation as a target variable, the role of the potential biases would have been overestimated, because they would have accounted for too much of the variance. Similarly, if we had collapsed all of the latent variables into a single average rating, the relationship between the potential biases would have been masked. It is only by separating out the different dimensions that a more precise picture of the relationships between the biases and the various evaluative dimensions can emerge.

We also examined the role of students' goals in how they reacted to their courses. We expected students whose primary motivation was to understand material would be more likely to enjoy courses than students whose primary goal was a good grade. However, the only achievement goal that emerged as important in our model was Mastery; the constructs of Performance-Approach and Performance-Avoidance did not figure significantly in any of the models we tested. Even Mastery, moreover, had only an indirect effect—Mastery affected Anticipation, which affected Involvement, which affected Overall Evaluation—and the effect was small.

Previous research using structural models has shown a positive relationship between mastery goals and enjoyment, and a negative relationship between performance-avoidance goals and enjoyment (Elliot & Harackiewicz, 1994, 1996; Elliot & Church, 1997; Elliot & McGregor, 2001; Harackiewicz *et al.*, 2000, 2002; Elliot & Thrash, 2002). One reason that these studies may have found achievement goals playing a more prominent role was that the dependent variables in these studies were related to measures of intrinsic motivation. Here, by contrast, a much wider range of measures was used to assess student reactions to their courses. Indeed, when we tested models in which the only dependent variables were questions involving interest and enjoyment, we also found that goals played a greater role. Our results, however, suggest that achievement goals may play a relatively small role in the kinds of course evaluation questionnaires used in most courses.

The role of missed expectations

As noted in the introduction, Remedios *et al.* (2000) found student reactions to their courses were influenced by the extent to which their actual grade differed from the grade they had expected. In our models, however, this variable proved to play no role. One possible explanation for this discrepancy is that Remedios *et al.* used only two measures of students' reactions, self-rated interest and enjoyment, whereas the current study used a much wider range of questions, and also included a much larger set of independent variables—for example, study hours and goal orientation. In this richer context, grade expectations did not emerge as a predictor in the best fitting

models. This is not to say that expectations do not play any role: We found that grades did have some influence on ratings, and some of the models that included expectations would have been valid if we had used more liberal criteria for goodness of fit. It would thus be premature to conclude that grade expectations do not play any role. Nevertheless, our modelling revealed stronger support for a validity hypothesis, and within this larger picture the effects of grade expectations appeared very small.

In other contexts, expectations might play a more important role. For example, Harackiewicz and Sansone's (1991) process model of intrinsic motivation identifies perceived competence at a forthcoming activity as an important mediator of intrinsic motivation. Structural modelling seems to provide a useful means by which the precise role that expectations play in determining experiences can be identified, and future research in this area might benefit from the use of this technique.

Conclusions

In summary, although grades did play some role in influencing overall ratings, the results from the study supported the validity hypothesis proposed by Marsh and Roche (2000), in that students' ratings of courses were largely determined by how well they felt they had been taught and how much they engaged with the course material. At least for our sample, ratings seemed to be earned rather than bought.

The results also revealed that a mastery-focus influenced how much students looked forward to their courses, and this in turn influenced overall ratings. In particular, students with a mastery orientation reported enjoying their courses more.

Our results also add weight to Marsh and Roche's (2000) conclusion that course evaluations need to be viewed as multidimensional. Figure 3 highlights how different characteristics of a course can be predicted by different variables. This points to students as discerning evaluators who are sensitive to different qualities of courses, and not individuals who if they rate one aspect highly are likely to rate all aspects highly. This conclusion should be reassuring to educators who organise their course carefully and to administrators who use evaluation instruments as measures of quality assurance. Factors such as grades, course difficulty and workload clearly do play a part in students' experiences and subsequent ratings of their courses, but our results suggest that students' ratings are better interpreted as measures of teaching quality rather than as rewards for good grades, low workload and non-challenging content.

Acknowledgements

This project was made possible by funding from the Economic and Social Research Council, award reference R000223608.

Notes

1. Since this study was first conducted, Elliot and McGregor (2001) have developed a fourth construct labelled mastery-avoidance. This construct is defined as a goal for avoiding missing opportunities to master a task.

2. Note that whilst it is usual to look for high values of χ^2 when examining differences, the opposite is the case when comparing models. This is because the best models are those that are similar to the optimal model, hence the phrases 'best-fitting models' and 'goodness-of-fit'.

References

Berger, J., Norman, R. Z. & Zeldich, M. J. (1977) *Status characteristics and social interaction* (New York, Elsevier).

Berger, J., Wagner, D. & Zeldich, M. J. (1985) Expectation states theory: review and assessment, in: J. Berger & M. J. Zeldich (Eds) *Status, rewards and influence* (San Francisco, CA, Jossey-Bass), 1–72.

Blunt, A. (1991) The effects of anonymity and manipulated grades on student ratings of instructors, *Community College Review*, 18, 48–54.

Bollen, K. A. (1989) *Structural equations with latent variables* (New York, Wiley).

Browne, M. W. & Cudek, R. (1993) Alternative ways of assessing model fit, in: K. Bollen & S. Long (Eds) *Testing structural equation models* (Newbury Park, NJ, Sage), 136–162.

Byrne, B. (2001) *Structural equation modelling with AMOS: basic concepts, applications and programming* (Mahwah, NJ, Lawrence Erlbaum).

Chako, T. I. (1983) Students' ratings of instruction: a function of grading standards, *Educational Research Quarterly*, 2, 341–351.

D'Apollonia, S. & Abrami, P. C. (1997) Navigating student ratings of instruction, *American Psychologist*, 52, 1198–1208.

Deci, E. L. (1975) *Intrinsic motivation* (New York, Plenum Press).

Deci, E. L. & Ryan, R. M. (1985) *Intrinsic motivation and self-determination in human behaviour* (New York, Plenum Press).

Dweck, C. S. & Leggett, E. L. (1988) A social cognitive approach to motivation and personality, *Psychological Review*, 95, 256–273.

Elliot, A. J. & Church, M. (1997) A hierarchical model of approach and avoidance achievement motivation, *Journal of Personality and Social Psychology*, 72, 218–232.

Elliot, A. J. & Harackiewicz, J. M. (1994) Goal setting, achievement orientation, and intrinsic motivation: a mediational analysis, *Journal of Personality and Social Psychology*, 66, 968–980.

Elliot, A. J. & Harackiewicz, J. M. (1996) Approach and avoidance achievement goals and intrinsic motivation: a mediational analysis, *Journal of Personality and Social Psychology*, 70, 461–475.

Elliot, A. J. & McGregor, H. A. (2001) A 2×2 achievement goal framework, *Journal of Personality and Social Psychology*, 80, 501–519.

Elliot, A. J. & Thrash, T. M. (2002) Approach-avoidance motivation in personality: approach and avoidance temperaments and goals, *Journal of Personality and Social Psychology*, 82, 804–818.

Elliott, E. S. & Dweck, C. (1988) Goals: an approach to motivation and achievement, *Journal of Personality and Social Psychology*, 54, 5–12.

Feather, N. T. (1961) The relationship of persistence at a task to expectation of success and achievement related motives, *Journal of Abnormal and Social Psychology*, 63, 552–561.

Feather, N. T. (1963a) The relationship of expectation of success to reported probability, task structure, and achievement related motivation, *Journal of Abnormal and Social Psychology*, 63, 231–238.

Feather, N. T. (1963b) The effect of differential failure on expectation of success, reported anxiety, and response uncertainty, *Journal of Personality*, 31, 289–312.

Feather, N. T. (1966) Effects of prior success and failure on expectations of success and subsequent performance, *Journal of Abnormal and Social Psychology*, 3, 287–298.

Feldman, K. A. (1976) Grades and college students' evaluations of their courses and teachers, *Research in Higher Education*, 4, 69–111.

Fischoff, B. (1975) Hindsight does not equal foresight: the effect of outcome knowledge on judgement under certainty, *Journal of Applied Social Psychology*, 18, 93–199.

Fischoff, B. & Beyth, R. (1975) 'I knew it would happen': remembered probabilities of once-future things, *Organisational Behaviour and Human Performance*, 13, 1–16.

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

Greenwald, A. G. & Gillmore, G. M. (1997b) No pain, no gain? The importance of measuring course workload in students' ratings of instruction, *Journal of Educational Psychology*, 89, 743–751.

Griffin, B. W. (2004) Grading leniency, grade discrepancy and student ratings of instruction, *Contemporary Educational Psychology*, 29, 410–425.

Harackiewicz, J. M. & Sansone, C. (1991) You can get there from here, in: M. L. Maeher & P. R. Pintrich (Eds) *Advances in motivation and achievement* (Greenwich, CT, JAI Press), 21–49.

Harackiewicz, J. M., Barron, K. E., Tauer, J. M. & Elliot, A. J. (2002) Predicting success in college: a longitudinal study of achievement goals and ability measures as predictors of interest and performance from freshman year through graduation, *Journal of Educational Psychology*, 94, 562–575.

Harackiewicz, J. M., Barron, K. E., Tauer, J. M., Carter, S. M. & Elliot, A. J. (2000) Short-term and long-term consequences of achievement goals predicting interest and performance over time, *Journal of Educational Psychology*, 92, 316–330.

Hoelter, J. W. (1983) The analysis of covariance structures: goodness-of-fit indices, *Sociological Methods and Research*, 11, 325–344.

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

Joreskog, K. G. & Sorbom, D. (1996) *LISREL 8: User's reference guide* (Chicago, IL, Scientific Software International).

MacCallum, R. C., Browne, M. W. & Sugawara, H. M. (1996) Power analysis and determination of sample size for covariance structure modelling, *Psychological Methods*, 1, 130–149.

Marsh, H. W. (1983) Multidimensional ratings of teaching effectiveness by students from different academic settings and their relation to student/course/instructor characteristics, *Journal of Educational Psychology*, 75, 150–166.

Marsh, H. W. (1987) Students' evaluations of university teaching: research findings, methodological issues, and directions for future research, *International Journal of Educational Research*, 1, 253–388.

Marsh, H. W. & Roche, L. A. (1997) Making students' evaluations of teaching effectiveness effective: the critical issues of validity, bias and utility, *American Psychologist*, 52, 1218–1225.

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

McKeachie, W. J. (1997) Student ratings: validity of use, *American Psychologist*, 52, 1218–1225.

Olivares, O. J. (2001) Student interest, grading leniency and teacher ratings: a conceptual analysis, *Contemporary Educational Psychology*, 26, 382–399.

Perry, R. P., Abrami, P. C., Leventhal, L. & Check, J. (1979) Instructor reputation: an expectancy relationship involving student ratings and achievement, *Journal of Educational Psychology*, 71, 776–787.

Pintrich, P. (2003) Multiple goals and multiple pathways in the development of motivation and self-regulated learning, *British Journal of Educational Psychology, BJEP Monograph series II*, 137–154.

Powell, R. W. (1977) Grades, learning and student evaluation of instruction, *Research in Higher Education*, 7, 193–205.

Remedios, R., Lieberman, D. A. & Benton, T. G. (2000) The effects of grades course enjoyment: did you get the grade you wanted?, *British Journal of Educational Psychology*, 70, 353–368.

Stringer, M. & Irwing, P. (1998) Students' evaluations of teaching effectiveness: a structural modelling approach, *British Journal of Educational Psychology*, 68, 409–426.

Stumpf, S. A. & Freedman, R. D. (1979) Expected grade covariation with student ratings of instruction: individual vs. class effects, *Journal of Educational Psychology*, 71, 293–302.

Tabachnick, B. G. & Fidell, L. S. (1996) *Using multivariate statistics* (3rd edn) (New York, Harper Collins).

Vasta, R. & Sarmiento, R. F. (1979) Liberal grading improves evaluations but not performance, *Journal of Educational Psychology*, 71, 207–211.

Worthington, A. G. & Wong, P. T. P. (1979) Effects of earned and assigned grades on student evaluations of an instructor, *Journal of Educational Psychology*, 71, 764–775.