MacEwan UNIVERSITY

RO@M Research Online at MacEwan

Easy Expectations and Racial Bias in Economics Instructor Ratings

Junaid B. Jahangir

NOTICE: This is the peer reviewed version of the following article: Jahangir, J. B. (2023). Easy expectations and racial bias in economics instructor ratings, Advances in Economics Education, 2(1), 90-107, which has been published in final form at http://dx.doi.org/10.4337/aee.2023.01.07.

Permanent link to this version https://hdl.handle.net/20.500.14078/3143

License All Rights Reserved

This document has been made available through RO@M (Research Online at Macewan), a service of MacEwan University Library. Please contact roam@macewan.ca for additional information.

"EASY EXPECTATIONS" AND RACIAL BIAS IN ECONOMICS INSTRUCTOR RATINGS

ABSTRACT

The objective in this paper is to investigate the determinants of Economics instructor ratings in two universities in Edmonton based on the data available from the Rate My Professors (RMP) website. Based on random effects and multi-level regression analysis, it is found that instructor ratings are predominantly driven by difficulty level and grades received by students. Additionally, ethnic instructors receive significantly lower ratings, which is explained less by accent and more by race. If the reported difficulty level of a course and the grade received by a student capture "easy expectations" on the part of students in the RMP data, and if instructor ratings are driven by a combination of such "easy expectations" and racial bias on the part of students, then the case for using average instructor ratings for annual faculty evaluations is weakened.

Keywords: Rate My Professors; racial bias; course difficulty; teaching Economics; teaching evaluations

"EASY EXPECTATIONS" AND RACIAL BIAS IN ECONOMICS INSTRUCTOR RATINGS

1. INTRODUCTION

In many cultures teaching has traditionally been viewed as a respected profession. The *guru* or *ustad* occupied a high position in the Indian subcontinent. Teaching often included training in punctuality, discipline and transfer of values from one generation to another. However, pedagogical methods that consisted of asking pupils to think on the spot, take notes dutifully, attend class on time, and write exams that tested critical thinking ability are perhaps considered "old school" and not rated favourably in teaching evaluations. Indeed, based on the popular Rate My Professors (RMP) website, the average overall ratings for retired University of Alberta professors is lower than those received by current contract instructors, who teach a bulk of the foundation level Economics courses (3.58/5 compared to 3.95/5).

However, a simple comparison of average overall ratings masks detailed information on teaching pedagogy and instructor attributes such as gender, ethnicity, and accent, all of which potentially impact an instructor's ratings. Poor ratings could be due to issues in teaching pedagogy or they may be dismissed as the ratings of students, who were expecting superior results based on minimum effort, easiness or "easy expectations". Easiness or "easy expectations" here refers to student perceptions that certain instructors reduce challenge or workload and liberally award As, thereby contributing to grade inflation through lower expectations of student performance. This phenomenon is noted

heavily in the literature, where several studies like Stark and Freishtat (2014) and Boring, Ottoboni and Stark (2016) have found high correlations between students' grade expectations and their evaluation of instructors. Moreover, such results incentivize instructors to maintain "easy expectations" by grading easier and setting easier assessments (Heffernan 2021).

Additionally, poor ratings could capture student bias as they rate women, ethnic minorities, and accented instructors more harshly compared to white male instructors who speak without a noticeable accent. Regardless, to the extent such unofficial RMP ratings are reflective of formal student evaluations conducted by educational institutions and to the extent such formal ratings are used in decisions to renew contracts for part time instructors or promotions for full time faculty members, it is important to highlight any biases that may arise in instructor ratings due to "easy expectations" or instructor attributes. Biases related to both instructor demographics including ethnicity, accent and gender along with student expectations of easy grades, have been effectively highlighted by a recent comprehensive literature review by Heffernan (2021).

The contribution of this paper lies in evaluating the presence of bias related to instructor ethnicity, accent and student expectation of ease in the context of teaching Economics at two universities in Edmonton. The focus is narrow, as it allows the author to observe accent, based on whether English is the first language of the instructor or whether an instructor whose first language is English has an accent that differs markedly from the local accent. Additionally, the focus is on Economics, as Economics instruction ratings

have been consistently ranked amongst the lowest due to factors including math based instruction. Generally, humanities and arts courses receive higher ratings than mathematics based courses, as many students feel incompetent in quantitative skills (Braskamp and Ory, 1994; Cashin, 1990; Neumann, 2000). Other reasons for low teaching evaluations in Economics include comparativey low grades awarded in Economics classes (Cashin, 1990).

The objective in this paper is to investigate the determinants of overall ratings of instruction quality of teaching Economics in two universities in Edmonton based on data available from the RMP website. The focus is on factors including difficulty level and received grades, which are related to "easy expectations" on behalf of students, along with instructor attributes on gender, ethnicity, and accent. The point of this study is to investigate whether student ratings of instructors are significantly driven by such factors for Economics instructors at two Edmonton based universities, MacEwan University and the University of Alberta. This is accomplished through a quantitative analysis of student ratings at the RMP website. Specifically, a random effects analysis, as used by McPherson, Jewell and Kim (2009) and to some extent the multi-level analysis approach of Baek and Shin (2008) is used in this paper.

The results of this study will allow us to question the suitability of using instructor ratings for annual faculty evaluations, salary raises, promotions and tenure decisions (Algozzine et al. 2004). In pursuit of this objective, this paper is divided into six sections. The next section delves into a select literature review on student ratings of instruction quality, justifying the use of data from the RMP website and biases in teaching evaluations.

The third section describes methodology on the data collected and the analytic strategy employed. The fourth section offers preliminary data analysis and the fifth section provides results from regression analysis. The final section offers concluding remarks on the results.

2. LITERATURE REVIEW

Student ratings of instruction quality

There is concern on whether students can evaluate instruction quality, as they are neither trained observers nor privy to instructor pedagogy (Braskamp et al. 1981). According to Becker (2000) there is low correlation between student evaluations and other measures of teaching effectiveness including tests scores and alumni surveys. However, Lattuca and Domagal-Goldman (2007) mention that a considerable body of research on teaching evaluations finds that students are good judges of clarity, preparation and organization but not content. Likewise, Feldman (2007) and Pan et al. (2009) repudiate claims that students lack maturity to evaluate instruction quality. According to Marsh (2007) teaching evaluations are helpful in improving instructional quality. On the other hand, it has been noted that teaching evaluations may simply reflect students' particular experience or disposition (MacFadyen, Dawson, Prest and Gasevic 2016) and primarily capture student satisfaction (Abrami et al. 2007; Beecham 2009). This means that while student ratings are supposed to measure instructional quality by rating factors like instructor clarity and organization, they may merely be capturing student satisfaction and experience.

Generally, according to Smith (2007), teaching effectiveness (in our context, quality) is gauged through student ratings on multidimensional items that include helpfulness, critical thinking, organization, workload, preparation and clarity. According to Shu-Hui and Goh (2003), the main determinants of student ratings include class size but neither instructor attributes like experience, rank and gender, nor course characteristics like subject type and level. However, McPherson et al. (2009) indicate that important determinants of student ratings include not only class size but also instructor experience, gender and instructor age. They also indicate that faculty may extract higher ratings by inflating student grade expectations (in our context, "easy expectations"), which is consistent with the findings of Weinberg et al. (2009) that student ratings are related to current grades but not learning. In short, student ratings of instructional quality is gauged through factors like organization, preparation and clarity, but bias in such ratings may creep through instructor attributes like gender and age or "easy expectations". It is in this general context of student ratings of instructional quality that the following discussion on RMP ratings and biases in teaching evaluations is situated and further reviewed.

Using the Rate My Professors data

Universal Student Ratings of Instruction (USRI) are officially used by educational institutions for teaching evaluations but the results are not publicly available. However, data on the RMP website are readily available and these ratings show strong correlations with formal university evaluations (Albrecht and Hoopes 2009; Timmerman 2008) and are consistent (Silva et al. 2008) with them. One reason for focusing on teaching evaluations from the RMP website is that students as customers rely on peer recommendations on

instructor and course selection (Harlow, 2003; Felton, Mitchell and Stinson, 2004). Another reason is the strong correlation between overall quality scores at the RMP website and the USRI teaching evaluation item "overall, how would you rate the instructor" (Coladarci and Kornfield, 2007). Otto and Sanford (2008) even argue that RMP ratings could be a useful supplement to teaching evaluations. While Legg and Wilson (2012) suggest that students on the RMP website have a negative bias and therefore are not representative of classes, Bleske-Rechek and Michels (2010) indicate that RMP ratings are moderate in tone as opposed to ranting and raving and tend to be more positive than negative. As such, while students who have a strong like or dislike for an instructor are more likely to leave ratings at the RMP website (Sen, Voia and Woolley, 2010) thereby biasing the results, this sample selection bias is also true for formal teaching evaluations. Therefore, using data from the RMP website does not pose an additional problem compared to formal teaching evaluations.

Biases in teaching evaluations

The literature indicates that while faculty is focused on learning, students may simply care for grades (Hornstein 2017) and these grades are significant determinants of teaching evaluations (Millea and Grimes 2002; Weinberg et al. 2009). Positive correlations are found between ratings of instructor quality and easiness (Rosen 2017) and student preference is noted for easy classes (Miller, 2006). That is, "easy expectations" play a role for some students in evaluations of faculty. However, Feldman (2007) notes that students who learn more and receive higher grades also may give higher instructor ratings. It is possible that easiness is capturing instructional quality, as Otto, Sanford and Ross (2008)

suggest that easiness could be interpreted as “easy to understand” instead of “not challenging.” Thus, there is ambiguity in whether better teachers are perceived as easier or whether easier teachers are perceived as better teachers (Rosen, 2017). According to Theyson (2015), the easiness and quality relationship may be complex as extremely hard and extremely easy may each be viewed as low quality.

Additionally, Heffernan (2021) notes in his literature review that academics may be motivated to set easier assessments and grade easier, especially if their livelihoods are at stake. Indeed, faculty members, especially contract instructors or those without tenure, who are concerned about potential contract renewal or promotion, would have the incentive and/or pressure to give easier exams, contribute to grade inflation and generally “dumb down” instructional material. There exists anecdotal evidence that reducing course challenge is often used strategically in an attempt to boost teaching evaluations (Trout 1997; 2000). Hornstein (2017) also notes that teaching evaluations pressurize faculty members to not rock the boat and to not push undergraduate students to maximize their intellectual potential.

While factors like easy grading and the intentional establishment of “easy expectations” are within the control of instructors, teaching evaluations are also affected by factors outside instructor control. An important finding is that racial minority faculty receive more negative ratings (Boatright-Horowitz and Soeung 2009; McPherson and Jewell 2007; Smith 2007). Alluding to the literature, Heffernan (2021) notes that the bias against instructor ethnicity is well documented. Smaller scale surveys and qualitative method studies including DiPietro and Faye (2005) and Hamermesh and Parker (2005) find

prejudice against academics of colour. Racial faculty may also resort to easy grading if upholding rigorous standards invites punishment through lower ratings (Clayson, Frost and Sheffet 2006). Likewise, lower ratings are associated with having an accent or not having English as the first language (Ogier 2005) just as if the instructor were more phenotypical of racial stereotypes (Dixon and Maddox 2005). Similarly, Finegan and Siegfried (2000) find that instructors for introductory Economics classes whose first language is not English received significantly lower ratings compared to instructors whose first language is English. More recently, Fan et al. (2019) find that academics from diverse backgrounds or those whose first language is not English receive lower ratings in student evaluations. This suggests that the impact of race on overall ratings has to be distinguished from that of accent to avoid confounding their effects.

In contrast to the literature that finds bias towards male faculty (Davison and Price 2009; Schmidt 2015), Anderson and Siegfried (1997) found no significant difference between the ratings of male and female instructors of introductory Economics classes. It has been noted that instructors of large introductory classes consistently receive lower scores than those of smaller advanced classes and elective courses (Pienta 2017). However, Davies et al. (2006) found no significant effect of class size on instructor ratings. Similarly, teaching experience and academic rank are not found to affect instructor ratings (Dynan and Rouse 1997; Shu-Hui and Goh 2003).

In terms of data analysis, multiple methods have been used apart from simple correlation analysis. While Reid (2010) used cluster analysis, Theyson (2015) used tobit analysis, as instructor quality data is presented via indicator variables that range from 1 - 5

and binary variables are used to capture instructor gender and hotness. Based on ols and logit regression analysis on RMP data, Boehmer and Wood (2017) tested the hypotheses on whether instructor ratings are higher for male faculty, easy instructors and from students with higher grades. However, for the purposes of this paper, random effects analysis, as used by McPherson, Jewell and Kim (2009) and to some extent the multi-level analysis approach of Baek and Shin (2008) is used. These approaches will be described in more detail in Sections 3 and 4 below.

In summary, the relationship between easiness and instructor quality is not clear, as easy instructors are rated highly by the subset of students who are focused on obtaining grades without expending effort, and as effective instructors can be perceived to be easy due to their clarity of instruction. Similarly, while there is some evidence of bias against instructor demographics including ethnicity, gender and accent, there is, some disagreement in the literature on gender. In terms of quantitative analysis, the RMP website data may not allow to address the relationship between easiness and instructor quality because of data limitations that preclude the use of instrumental variables to account for this endogeneity. Similarly, it only allows us to test for the impact of received grades as opposed to expected grades. However, to the extent that students provide the correct course code, the RMP data allow us to test for the impact of course level on overall ratings. Additionally, using data on both ethnicity and accent allows us to discern if instructor ratings were motivated by discomfort in understanding foreign accents or by racial bias.

3. DATA AND METHODS

MacEwan University and the University of Alberta in Edmonton provide an excellent case to quantitatively investigate the determinants of Economics instructor ratings and to discern any bias related to ethnicity, gender and accent. This is because of the close connection between the two. Where some faculty at MacEwan have been trained at the University of Alberta, many students have historically transferred course credits from one institution to the other. Additionally, Economics courses at MacEwan University, a teaching-intensive institution, are parallel to those taught at the University of Alberta, a research-intensive institution, where both tenure track and contract-based teaching faculty teach Economics students. This facilitates comparison of teaching evaluations between tenure track faculty and contract-based lecturers. The two institutions, however, differ due to larger class sizes and the relatively stringent entrance requirements at the University of Alberta. A distinguishing feature for MacEwan is that Economics is part of a portmanteau department of Anthropology, Economics and Political Science (AEPS), where the same broad governance structure, including common department meetings, deadlines and general expectations, also allows for comparing instructor ratings across the three disciplines.

The RMP website provides ratings on 83 instructors in terms of six distinct categories. Table 1 shows the eight questions that students answer in order to rate the instructor. The chief question is on instructor quality and is based on the question "How would you rate this professor as an instructor?" This question is answered on a 1 – 5 scale, which is associated with three emoticons - green for awesome (4.0, 5.0), yellow for average (3.0) and red for awful (1.0, 2.0). The question on difficulty level is also answered on a 1-

5 scale, which gauges the range between an easy A and working hard. The other four questions are answered in a binary fashion and include whether the student would take another class with the instructor, if the course was taken for credit, whether the textbook was used and if attendance was mandatory. Additionally, there are two questions that pertain to the course code and the grade received by the student. Finally, there is space for a comment that is based up to a maximum of 350 characters.

According to the RMP site guidelines, ratings can only be given by former or current students and they are limited to only one rating per course. Usually students rate professors after having taken the course or during the course, however, on some occasions, an older student is noted to have left a comment, as in the case of a retired instructor that passed away in the current sample. The site is moderated for violations of guidelines including spamming and remains the largest platform for potential students to review instructor ratings.

In terms of the grade distribution of the current sample, of the 539 student ratings, more than 80% self-reported a grade of B+ or higher and about 52% reported a grade of A. This is consistent with the findings in the literature noted above that grades are a significant determinant of the submission of teaching evaluations. Moreover, in terms of quality assurance of data, as noted through the above literature review, while students who have a strong like or dislike for an instructor are more likely to leave ratings at the RMP website thereby biasing the results, this sample selection bias is also true for formal teaching evaluations. Therefore, using data from the RMP website does not pose an additional problem compared to formal teaching evaluations.

The RMP student ratings include instructors in Anthropology, Political Science and Economics at MacEwan University and full time, retired and contract Economics instructors at the University of Alberta in Edmonton. These ratings provide instructor evaluations based on six student/course level variables including instructor difficulty level, percentage of students that would take another class with the instructor, the course level, whether students used the textbook, whether students deemed attendance as mandatory, and the grade received. Data points for these variables were collected for approximately a five-year period from 2015 – 2019. Student ratings from 2020 are ignored to avoid the impact of COVID-19 due to a shift towards online instruction. Data on instructor demographics including ethnicity, accent and gender are based on the author's first-hand knowledge based on working at both institutions, as the instructors in the current sample are either former teachers or colleagues of the author with interaction through department meetings or other social events. In terms of the 539 student ratings in the sample, about 38% are for accented instructors with the majority 62% for native English speakers. In short, student ratings seem to be driven by students who have received high grades and for a majority of native English speaking instructors. This means that student ratings would be expected to be skewed positively towards high ratings, and therefore econometric modeling would help to identify any bias pertaining to ethnicity or accent.

As mentioned earlier, the focus in this paper is narrow as it focuses on Economics instructors at two universities in Edmonton. This allows the author to gauge instructor ethnicity and accent. While individual self-reported ethnicity could differ, for instance a person with Caucasian features could reject being labelled as "white" based on rejection of that classification, for the purposes of this paper, instructors were noted as white based on

Caucasian features, which included North Americans, East and West Europeans, Australian and Russians. On the other hand, most instructors from Asian, South American and African heritage were noted as people of colour. In terms of accent, the decision was made to consider native English speakers as non-accented and non-native English speakers as accented. The latter includes East Europeans and Russians apart from Asians, Africans and South Americans. Exceptions were made in the case where the instructor had a clear North American style of speaking, as in the case of an instructor of East European heritage, one of Asian heritage and another of Indian heritage.

Nevertheless, given that there can be differences in perceptions of ethnicity and accent, feedback from others suggested alternative categorizations for a small minority of the faculty in the sample based on geographical variations in accents for native English speakers and similar countries of origin for instructors who may be perceived by some but not all as belonging to racial or ethnic minorities. Sensitivity results with alternative classifications for some instructors did not change the results in Section 5.

Data on the percentage of white instructors, native English speakers and male instructors along with the average instructor quality in each of the six instructor categories are summarized in Table 2. This Table distinguishes students in 100 level courses from others and provides the average instructor difficulty level and the average grade received for each of the six instructor categories. Table 2 also provides this data by distinguishing MacEwan instructors between continuing and contract instructors. For quantitative analysis, data based on Table 2 provides observations on the six student/course level

variables along with instructor quality and instructor demographics. Overall, 539 observations are available for regression analysis.

Since the data provides information on 83 instructors for a span of about five years, failure to account for instructor heterogeneity leads to biased results and necessitates panel data modeling. Additionally, failure to account for statistical dependency among observations, as students who take a particular class may be more similar to each other, leads to biased standard errors and necessitates multilevel modeling. Data is modeled across instructor and time for panel data analysis. For multilevel modeling, students at level one are nested within instructors at level two. Random effects estimation allows incorporating time invariant variables like instructor gender and ethnicity. It may be noted that a random effects model is a simple hierarchical linear model with a random intercept, and therefore a special case of multilevel analysis. Overall, an unbalanced panel data model of the determinants of instruction quality, based on instructor demographics and the six student level variables, is used for analysis.

4. PRELIMINARY DATA ANALYSIS

Table 2 indicates that MacEwan Economics faculty members receive the lowest quality ratings from students. It suggests that this may be due to four factors – the lowest average grade received by students (3.43/4), ratings by predominantly students at the first-year level (80.23%) and the highest percentage of ethnic and accented faculty members (72.73%). Table 1 also indicates that the difficulty rating of retired Economics faculty members at the University of Alberta is the highest (3.32/5). Additionally, class attendance and the

willingness to take a class again with the instructor is much higher for Anthropology and Political Science instructors compared to those in Economics.

Distinguishing MacEwan faculty between continuing and contract instructors, Table 2 indicates that the difficulty level of contract instructors, whether at MacEwan (2.48/5) or at the University of Alberta (2.69/5), is the lowest and the quality rating is higher compared to continuing faculty members in Economics (3.69 versus 3.26 for MacEwan and 3.95 versus 3.50 for the University of Alberta). Additionally, the average grade received by students is also higher for contract instructors compared to continuing instructors in Economics (3.55 versus 3.41 for MacEwan and 3.72 versus 3.64 for the University of Alberta). This suggests either that contract instructors have lowered academic standards through "easy expectations" and easy grades or that their exposition is clearer compared to continuing Economics instructors.

In summary, the preliminary analysis shows that low ratings of Economics instructors are driven by first year level students, several of whom are required to take the class, and who have received lower grades. It also shows that ethnic instructors that speak with an accent receive the lowest quality ratings. Finally, for contract instructors, while the difficulty levels are lower and both instructor quality and grades received by students are higher, it is not clear if this is due to clearer exposition or lowering of academic standards.

5. REGRESSION ANALYSIS

The objective in undertaking regression analysis is to investigate the determinants of Economics instructor ratings and to discern any bias related to ethnicity, gender and accent.

In this regard, a model of the determinants of instruction quality was developed based on data availability of instructor demographics and student level variables. Overall, 539 data observations are available for regression analysis. The model uses instruction quality as the dependent variable and both student level variables and instructor demographics as explanatory variables. Following the notation in Worrall (2010), this panel data model estimated through random effects can be symbolically presented as follows.

$$Quality_{it} = \alpha + \beta X_{it} + \gamma Z_{it} + u_i + w_t + \varepsilon_{it}$$

The dependent variable on instructor quality (denoted by $Quality_{it}$) ranges between 1 and 5 and is based on the RMP question asked of the students that, "How would you rate this professor as an instructor?" The eight demographic variables for the instructors (denoted by X_{it}) are dummy variables including institution type, subject type (where Econ =1), ethnicity (where white = 1), accent (where native English accent =1), full-time status, gender (where male = 1), provision of notes and retired status. The variable on difficulty level ranges between 1 and 5, grades range between 0 and 4, whereas others are dummy variables. While there are six student level variables (denoted by Z_{it}), the variable on taking another class with the instructor is dropped, as it is simply an alternate measure of instructor quality. The five student level variables include difficulty level, class level, attendance, textbook requirement and grade received. For random effects panel data analysis, u_i and w_t are error terms that refer to the instructor specific effect and the time period effect respectively, whereas ε_{it} is a normally distributed error term. The summary statistics of the model variables are presented in Table 3. Based on this model, the idea is to test whether

instructor characteristics like ethnicity and accent and whether easiness or "easy expectations", which may be captured to some extent by difficulty level and grades received, are significant determinants of instructor quality.

Initial diagnostic testing of an OLS regression that does not account for heterogeneity or statistical dependency between observations and leads to biased results indicated the presence of heteroskedasticity, based on the Breusch Pagan test and functional form problems, based on the Ramsey RESET test. The Breusch Pagan LM test indicated presence of random effects, and the Hausman test favoured the random effects model instead of fixed effects. Using robust standard errors, the random effects model indicated the statistical significance of difficulty level, with a negative sign, and grades, with a positive sign, and since it allows for time invariant variables, it indicates the statistical significance for white instructors, with a positive coefficient, and male instructors, with a negative sign. The difficulty level and grade variables are signficant at 1%, the dummy variable for white instructors at 5% and the dummy variable for male instructors at 10% significance level. Additionally, the premium for white instructors is about 0.58 points more, which is sizeable as quality ranges between 1 and 5. This model was run again to control for instructor quality by creating a dummy variable 'Prof', which is equal to 1 for Associate and Full Professors and 0 otherwise. The third column in Table 4 indicates that the impact of this variable is negative, which suggests that more experienced instructors are rated lower. However, this variable is statistically insignificant and therefore it is dropped for subsequent analysis.

In order to check for robustness of results to the estimation method, a logit model with random effects was estimated where the dependent variable assumed the value of 1 for instructors rated higher than 3.5 and 0 otherwise. The results confirm the results of the random effects model for the statistical significance of difficulty level, grades and ethnicity. This logit model was run again where the dependent variable assumed the value of 1 for instructors rated higher than 4.0 and 0 otherwise. The results in column 7 of Table 4 confirm the results of the random effects model for the statistical significance of difficulty level, grades and gender but not ethnicity. This suggests that for very highly rated instructors, neither ethnicity nor accent offers any premium. Another test was undertaken to check for the robustness of results. Essentially, since both difficulty level and grades could have potential collinearity, in that grades may be lower if the difficulty level is higher, each of these variables were dropped and the overall random effects model was estimated. The findings in columns 4 and 5 of Table 4 confirm the results on the statistical significance of difficulty level, grades and ethnicity. Yet another sensitivity analysis included the classification of a couple of instructors on ethnicity and accent. Based on input from three colleagues, who differed on the ethnicity of one instructor and accent of another, the analysis was repeated but the impact of alternative classification was negligible, Thus, the results are also robust to a slight change in the classification of instructors.

Modeling the data through a multilevel model allowed another check for robustness to the estimation method. As noted earlier, failure to account for statistical dependency among observations, as students who take a particular class may be more similar to each other, leads to biased standard errors and necessitates multilevel modeling. Based on Baek and Shin (2008), if student response data from RMP were used as the unit of analysis, the

independence assumption would be violated and the variance is inflated. This necessitates multilevel modeling, where students at level one are nested within instructors at level two. It may be noted that a random effects model is a simple hierarchical linear model with a random intercept, and therefore a special case of multilevel analysis. Following Baek and Shin (2008), this multilevel model, where student level variables (Level 1) are nested in instructor level variables (Level 2), can be symbolically presented as follows.

$$\begin{aligned} Quality_{ij} = {} & \gamma_{00} + \gamma_{10}\left(Diff - \overline{Diff}\right) + \gamma_{20}\,100\,level + \gamma_{30}\,Attendance \\ & + \gamma_{40}\,Text + \gamma_{50}\left(Grade - \overline{Grade}\right) + \gamma_{01}\,Econ + \gamma_{02}\,Uni \\ & + \gamma_{03}\,Ethnic + \gamma_{04}\,English + \gamma_{05}\,Full\ time + \gamma_{06}\,Gender \\ & + \gamma_{07}\,Notes + \gamma_{08}\,Retired + u_{0j} + r_{ij} \end{aligned}$$

In this equation, $Quality_{ij}$ refers to the student rating i received by instructor j, γ_{00} denotes the grand mean, which indicates the average instructor quality across all instructors, u_{0j} denotes the difference of instructor j's rating from the grand mean, and r_{ij} denotes the difference of instructor j's rating from the average of instructor j's ratings. In contrast to OLS regression, where the intercept and coefficients are fixed effects, and the residual term is a random effect, in multilevel modeling, there are several intercepts and coefficients for each unit (in our case, instructor) at Level 2. Additionally, the intercepts in a multilevel model are estimated as a random effect, as they vary between Level 2 units (instructors). This variation between Level 2 units (instructors) account for the non-independence between Level 1 units (students). The analysis starts with an unconditional model without any predictors, where Level 2 units (instructors) differ from each other on the outcome variable (instructor quality). Then explanatory variables are added to that

unconditional model. Explanatory variables, difficulty level and grades are grand mean centred by subtracting the sample mean from each score on these variables. This allows for a useful interpretation of the grand mean γ_{00}, in that it would be viewed as the mean instructor quality for instructors rated at average difficulty level and who awarded the average grade to students, assuming 0 values for all the other dummy variables in the model.

The unconditional model with just an intercept indicates the interclass correlation coefficient (ICC) of 0.445, which indicates that the proportion of variance explained by instructor heterogeneity is 44.5%. Adding explanatory variables to the multilevel model, where difficulty level and grades are grand mean centred, decreased the ICC to 0.355. Finally, when random coefficients for difficulty level and grades are allowed, the ICC becomes 0.357. The Likelihood ratio test favours the multilevel model with random coefficients. However, for both random intercepts and random coefficients, the statistical significance of difficulty level, grades and ethnicity holds. Table 4 summarizes the results of the random effects model, logit model with random effects and multilevel model with random coefficients. The results indicate that the significance of difficulty level, grades and ethnicity is robust to the estimation method.

Overall, the results indicate that instructor ratings are driven by student concerns on difficulty level and grades received. However, data limitations preclude addressing the potential endogeneity issue due to difficulty and grades, as quality ratings can influence the instructor to modify difficulty level and grades. This limitation precludes addressing whether better teachers are perceived as easier or whether easier teachers are perceived as

better teachers (Rosen 2017). To get a better picture of how "easy expectations" affect student ratings of instructors, instrumental variable (IV) methods will have to be used, which require data on variables that could be used as potential instruments.

The statistical significance of ethnicity but insignificance of accent suggests that it is not discomfort in understanding foreign accents that is behind the lower ratings of ethnic instructors. This result is different in the context of the literature that shows lower ratings for both racial minorities and for having an accent. Finally, the random effects panel data model indicates that male instructors receive statistically significantly lower ratings, which confirms the observation that the average male instructor rating is 3.66 compared to 3.85 for female instructors for the entire data set. This result stands in contrast to the literature that finds bias towards male faculty (Davison and Price 2009; Schmidt 2015). The results also indicate that class level, instructor status as continuing or contract, and the provision of notes are insignificant in determining instructor quality. This result stands in contrast to the literature that finds that large introductory classes consistently receive lower scores (Pienta 2017) but confirms the findings in the literature that shows that academic rank does not affect instructor ratings.

6. CONCLUSIONS

The objective in this paper was to investigate the determinants of Economics instructor ratings, including the impact of "easy expectations" and to discern any bias related to ethnicity, gender and accent. Regression analysis including random effects panel data analysis and multilevel modeling indicate that lower difficulty level and higher grades received are significant determinants of higher instructor quality. However, while lower

difficulty level and higher grades received are associated with contract instructors, it is not clear if this is due to better teaching pedagogy or "easy expectations" and lowering of academic standards. The result on male instructors receiving statistically significantly lower ratings was not robust to differences in estimation methods. On the other hand, while the variable on accent was insignificant, the result on white instructors receiving higher ratings was statistically significant and robust to the estimation techniques used, although it became statistically insignificant for very highly rated instructors in the logit model with random effects. This suggests that the ratings of ethnic instructors in the RMP data, based on two universities in Edmonton, are not explained by discomfort in understanding foreign accents and that there is no statistically significant difference in the ratings of very highly rated white and ethnic instructors.

This study contributes to the literature by focusing specifically on Economics education. It shows that ratings of Economics instructors suffer from the same biases related to course difficulty, possibly attributable to "easy expectations" and racial bias, as has been generally found in student ratings across academic disciplines. The focus on Economics is specially warranted because even though the subject is part of social sciences, it is predominantly a quantitative subject, and therefore, instructor ratings are lower than those of other disciplines in the social sciences. The findings of this small scale study add a finer perspective to the literature that any racial bias is not due to the accent or English not being the first language of the instructor. Similarly, it adds a different perspective that male instructors are rated lower than their female counterparts.

However, any concrete conclusion on "easy expectations" and racial bias in instructor ratings is limited based on the limitations of the methodology used, and also due to the fact that RPM ratings do not include questions on teaching style and pedagogy. Nonetheless, if difficulty level and grades received are capturing "easy expectations" and to the extent instructor ratings are driven by "easy expectations" and racial bias, and to the extent RMP ratings are consistent with formal University instructor ratings, the case for solely using average instructor ratings for annual faculty evaluations, salary raises, promotions and tenure decisions is weakened. The results obtained from the RPM data suggest that these issues warrant further investigation for not only Economics instruction but in any field where a diverse set of instructors are tasked with educating the next generation of scholars and skilled workers.

References

Abrami, P. C., d'Apollonia, S. and Rosenfield, S. (2007) 'The dimensionality of student ratings of instruction: What we know and what we do not', in Perry, R. P. and Smart, J. C. (eds.) *The scholarship of teaching and learning in higher education: An evidence-based perspective,* Springer, New York, pp. 385–456.

Albrecht, S. and J. Hoopes. 2009. 'An empirical assessment of commercial web-based professor evaluation services', *Journal of Accounting Education,* 27(3), pp. 125-132.

Algozzine, B., Gretes, J., Flowers, C., Howley, L., Beattie, J., Spooner, F., Mohanty, G. and Bray. M. (2004) 'Student Evaluation of College Teaching: A Practice in Search of Principles', *College Teaching,* 52(4), pp. 134−141.

Anderson, H. and Siegfried, J.J. (1997) 'Gender Differences in Rating the Teaching of Economics', *Eastern Economic Journal,* 23(3), pp. 347-357.

Baek, S.G and Shin, H.J. (2008) 'Multilevel analysis of the effects of student and course characteristics on satisfaction in undergraduate liberal arts courses', *Asia Pacific Education Review,* 9(4), pp. 475-486.

Becker, W.E. (2000) 'Teaching Economics in the 21st century', *The Journal of Economics Perspectives,* 14 (1), pp. 109-119.

Beecham, R. (2009) 'Teaching quality and student satisfaction: Nexus or simulacrum?' *London Review of Education,* 7, pp. 135–146.

Bleske-Rechek, A. and Michels, K. (2010) 'RateMyProfessors.com: Testing Assumptions about Student Use and Misuse,' *Practical Assessment, Research & Evaluation,* 15(5), pp. 1-12.

Boatright-Horowitz, S. and Soeung, S. (2009) 'Teaching white privilege to white students can mean saying good-bye to positive student evaluations', *American Psychologist,* 64(6), pp. 574–575.

Boehmer, D. M. and Wood, W. C. (2017) 'Student vs. faculty perspectives on quality instruction: Gender bias, "hotness," and "easiness" in evaluating teaching', *Journal of Education for Business*, 92(4), pp. 173-178.

Boring, A., Ottoboni, K., and Stark, O. (2016) 'Student evaluations of teaching (mostly) do not measure teaching effectiveness', *Science Open Research*, https://www.scienceopen.com/hosted-document?doi=10.14293/S2199-1006.1.SOR-EDU.AETBZC.v1 (Accessed 16 November 2021)

Braskamp, L., Ory, J. and Pieper, D. (1981) 'Student written comments: Dimensions of instructional quality', *Journal of Educational Psychology,* 73(1), pp. 65-70.

Braskamp, L. A. and Ory, J.C. (1994) *Assessing faculty work: enhancing individual and institutional performance,* Jossey-Bass, San Francisco.

Cashin, W. (1990) 'Students do rate different academic fields differently,' in Theall, M. and Franklin, J., (eds.) Student ratings of instruction: Issues for improving practice, New Directions for Teaching and Learning 43 (Fall), Jossey-Bass, San Francisco, pp. 113-21.

Clayson, D. E., Frost, T.F., and Sheffet, M.J. (2006) 'Grades and the student evaluation of instruction: A test of the reciprocity effect', *Academy of Management Learning & Education*, 5(1), pp. 52–85.

Coladarci, T. and Kornfield, I. (2007) 'RateMyProfessors.com versus formal in-class student evaluations of teaching,' *Practical Assessment, Research & Evaluation,* 12(6), pp. 1- 15.

Davies, W.M., Hirschberg, J., Lye, J., Johnston, C. and McDonald, I. (2006) 'What

Influences Teaching Evaluations? Evidence from a Major Australian University', *The Business Review, Cambridge,* 6(1), pp. 146-152.

Davison, E. and Price, J. (2009) 'How do we rate? An evaluation of online student evaluations,' *Assessment & Evaluation in Higher Education,* 34(1), pp. 51-65.

DiPietro, M. and Faye, A. (2005) 'Online student ratings of instruction (SRI) mechansims for maximal feedback to instructors', 30th Annual meeting of the professional and organizational development network, Milwaukee, Wisconsin.

Dixon, T. and Maddox, K.B. (2005) 'Skin tone, crime news, and social reality judgments: Priming the stereotype of the dark and dangerous black criminal', *Journal of Applied Social Psychology,* 35(8), pp. 1555–1570.

Dynan, K. E., and Rouse, C.E. (1997) 'The Under Representation of Women in Economics: A Study of Undergraduate Economics Students', *Journal of Economic Education,* 28(4), pp. 350-368.

Fan, Y., Shepherd, L., Slavich, D., Waters, D., Stone, M., Abel, R., and Johnston, E. (2019) 'Gender and cultural bias in student evaluations: why representation matters', Plos ONE, 14(2), https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0209749 (Accessed 16 November 2021)

Feldman, K. A. (2007) 'Identifying exemplary teachers and teaching: Evidence from student ratings,' in Perry, R. P. and Smart, J. C. (eds.) *The scholarship of teaching and learning in higher education: An evidence-based perspective,* Springer, New York, pp. 93–143.

Felton, J., Mitchell, J. and Stinson, M. (2004) 'Web-based student evaluations of professors: the relations between perceived quality, easiness and sexiness', *Assessment & Evaluation in in Higher Education,* 29(1), pp. 91-108.

Finegan, A. T., and Siegfried, J.J. (2000) 'Are Student Ratings of Teaching Effectiveness Influenced by Instructors' English Language Proficiency?' *American Economist*, 42(2), pp. 34-46.

Hamermesh, D., and Parker, A. (2005) 'Beauty in the Classroom: Instructors' Pulchritude and Putative Pedagogical Productivity', *Economics of Education Review,* 24(4), pp. 369-376.

Harlow, R. (2003) '"Race doesn't matter, but …": The effect of race on professors' experiences and emotion management in the undergraduate college classroom', *Social Psychology Quarterly,* 66(4), pp. 348–363.

Heffernan, T. (2021) 'Sexism, racism, prejudice, and bias: a literature review and synthesis of research surrounding student evaluations of courses and teaching', *Assessment & Evaluation in Higher Learning,* https://www.tandfonline.com/doi/abs/10.1080/02602938.2021.1888075. (Accessed 16 November 2021)

Hornstein, H.A. (2017) 'Student evaluations of teaching are an inadequate assessment tool for evaluating faculty performance', *Cogent Education,* 4(1), 1304016.

Lattuca, L.R. and Domagal-Goldman, J.M. (2007) 'Using qualitative methods to access teaching effectiveness', *New Directions for Institutional Research,* 136, pp. 81-93.

Legg, A.M. and Wilson, J.H. (2012) 'RateMyProfessors.com Offers Biased Evaluations', *Assessment & Evaluation in Higher Education,* 37(1), pp. 89-97.

Macfadyen, L.P., Dawson, S., Prest, S. and Gasevic, D. (2016) 'Whose feedback? A multilevel analysis of student completion of end of term teaching evaluations', *Assessment and Evaluation in Higher Learning,* 41(6), pp. 821-839.

Marsh, H. W. (2007) 'Students' Evaluations of University Teaching: Dimensionality, Reliability, Validity, Potential Biases and Usefulness', in Perry, R. P. and Smart, J. C. (eds.) *The Scholarship of Teaching and Learning in Higher Education: An Evidence-Based Perspective*, Springer, New York, pp. 319-383.

McPherson, M. A. and Jewell, R.T. (2007) 'Leveling the playing field: Should student evaluation scores be adjusted?' *Social Science Quarterly,* 88(3), pp. 868–881.

McPherson, M. A., Jewell, R.T. and Kim, M. (2009) 'What determines student evaluation scores? A random effects analysis of undergraduate Economics classes', *Eastern Economic Journal,* 35(1), pp. 37-51.

Millea, M. and Grimes, P.W. (2002) 'Grade expectations and student evaluation of teaching', *College Student Journal,* 36(4), pp. 582–590.

Miller, J.D. (2006) 'How to fight ratemyprofessors.com', *Inside Higher Ed*, January 31.

Neumann, R. (2000) 'Communicating student evaluation of teaching results: rating interpretation guides', *Assesment & Evaluation in Higher Education,* 25(2), pp. 121-134.

Niu, Shun-Chen. (2005) 'Comparing Two Populations', December 22, University of Texas at Dallas. https://www.utdallas.edu/~scniu/OPRE-6301/documents/Two_Populations.pdf (Accessed 26 March 2021)

Ogier, J. (2005) 'Evaluating the effect of a lecturer's language background on a student rating of teaching form', *Assessment & Evaluation in Higher Education,* 30(5), pp. 477–488.

Otto, J., Sanford, D.A.J. and Ross, D.N. (2008) 'Does ratemyprofessor.com really rate my professor?' *Assessment & Evaluation in Higher Education,* 33(4), pp. 355-368.

Pan, D., Tan, G., Ragupathi, K., Booluck, K., Roop, R. and Ip, Y. (2009) 'Profiling teacher/teaching using descriptors derived from qualitative feedback: Formative and summative applications', *Research in Higher Education,* 50(1), pp. 73-100.

Pienta, N. J. (2017) 'The slippery slope of student evaluations', *Journal of Chemical Education,* 94(2), pp. 131-132.

Reid, L. D. (2010) 'The role of perceived face and gender in the evaluation of college teaching on ratemyprofessors.com', *Journal of Diversity in Higher Education,* Vol. 3 No. 3, pp. 137-152.

Rosen, A. S. (2017) 'Correlations, trends and potential biases among publicly accessible web-based student evaluations of teaching: a large-scale study of ratemyprofessors.com data', *Assessment and Evaluation in Higher Education,* 43(1), pp. 31-44.

Schmidt, B. (2015) 'Gender bias exists in professor evaluations', *The New York Times*, December 16.

Sen, A., Voia, M. and Woolley, F. (2010) 'Hot or Not: How appearance affects earnings and productivity in academia', *Carleton Economic Papers* 10-07.

Silva, K. M., Silva, F.J., Quinn, M.A., Draper, J.N., Cover, K.R. and Munoff, A. A. (2008) 'Rate my professor: Online evaluations of psychology instructors', *Teaching of Psychology,* 35(2), pp. 71–80.

Shu-Hui, L. and Goh, K. (2003) 'Evidence and Control of Biases in Student Evaluations of Teaching', *The International Journal of Educational Management,* 17(1), pp. 37- 43.

Smith, B. P. (2007) 'Student ratings of teacher effectiveness: An analysis of end-of-course faculty evaluations', *College Student Journal,* 41(4), pp. 788–800.

Stark, P. and Freishtat, R. (2014) 'An evaluation of course evaluations', *Science Open Research*, https://www.scienceopen.com/hosted-document?doi=10.14293/S2199-1006.1.SOR-EDU.AOFRQA.v1 (Accessed 16 November 2021)

Theyson, K. C. (2015) 'Hot or not: The role of instructor quality and gender on the formation of positive illusions among students using ratemyprofessors.com', *Practical Assessment, Research & Evaluation,* 20(4), pp. 1 -12.

Timmerman, T. (2008) 'On the Validity of RateMyProfessors.com', *The Journal of Education for Business,* 84(1), pp. 55-61.

Trout, P. (1997) 'How to improve your teaching evaluation scores without improving your teaching!' *The Montana Professor,* 7(3), pp. 17–22.

Trout, P. (2000) 'Teacher evaluations', *Commonweal,* 127(8), pp. 10–11.

Weinberg, B. A., Fleisher, B. M. and Mashimoto, M. (2009) 'Evaluating Teaching in Higher Education', *The Journal of Economic Education,* 40(3), pp. 227-261.

Worrall, J.L. (2010) 'A user-friendly introduction to panel data modeling', *Journal of Criminal Justice Education*, 21(2), pp. 182-196.

Table 1: Rate My Professors questions

Question	Response
Course Code	Selection of course
How would you rate this professor as an instructor?	1 – 5
How hard did you have to work for this class?	1 – 5 (the scale on easy A versus working hard)
Would you take this Prof again?	Yeah - Um, No
Was this class taken for credit?	Yeah - Um, No
Textbook use	Yeah - Um, No
Attendance	Mandatory – Non Mandatory
Grade Received	Selection of grade
Here's your chance to be more specific	Comment (350 characters) The site gives suggestions as: your unique experience writing / reading intensity attendance policy availability outside of class required participation

Table 2: Summary of the Rate My Professors data

Instructors	# of instructors	White	Native English	Male	Quality	Difficulty	Would take again (%)	100 level	Attendance	Text used	Average grade
MacEwan Economics	11	27.27%	27.27%	100.00%	3.46	2.71	62.87%	80.23%	49.33%	64.86%	3.43
MacEwan Political Science	9	66.67%	55.56%	77.78%	3.92	2.92	78.04%	64.29%	68.95%	44.33%	3.61
MacEwan Anthropology	14	85.71%	85.71%	35.71%	3.86	2.75	79.86%	65.86%	56.57%	70.57%	3.62
UfA Full Time	22	68.18%	36.36%	59.09%	3.50	3.10	63.05%	1.43%	50.48%	44.70%	3.64
UfA Retired	14	92.86%	92.86%	85.71%	3.58	3.32		7.51%		42.12%	
UfA Contract Teaching	13	69.23%	53.85%	84.62%	3.95	2.69	65.30%	32.87%	47.59%	42.62%	3.72
Sum	83										
MacEwan Economics Continuing	7	14.29%	0.00%	100.00%	3.26	2.87	55.72%	68.94%	48.53%	52.53%	3.41
MacEwan Political Science Continuing	7	71.43%	57.14%	85.71%	3.97	3.09	80.85%	60.12%	72.07%	40.90%	3.58
MacEwan Anthropology Continuing	10	90.00%	90.00%	30.00%	4.00	2.81	79.49%	54.43%	58.84%	64.80%	3.65
MacEwan AEPS Contract	10	60.00%	70.00%	70.00%	3.69	2.48	76.08%	92.40%	52.48%	79.26%	3.55

**Empty cells indicate missing data*

Table 3: Summary Statistics and Variable Definition

Variable	Definition	Mean	Variance	Minimum	Maximum
Quality	Instructor overall rating with range 1 – 5	4.15	1.43	1	5
Logit	Dummy = 1 for instructor rated more than 3.5, 0 otherwise	0.77	0.18	0	1
Diff	Instructor Difficulty level with range 1 - 5	2.70	1.25	1	5
100 level	Dummy = 1 for 100 level classes, 0 otherwise	0.70	0.21	0	1
Attendance	Dummy = 1 if attendance is mandatory, 0 otherwise	0.57	0.25	0	1
Text	Dummy = 1 if text is used, 0 otherwise	0.53	0.25	0	1
Grade	Grade received by student from 0 – 4	3.63	0.31	0	4
Econ	Dummy = 1 for Econ, 0 for Political Science and Anthropology	0.60	0.24	0	1
Uni	Dummy = 1 for University of Alberta, 0 for MacEwan instructor	0.36	0.23	0	1
White	Dummy = 1 for white, 0 for ethnic minority instructor	0.64	0.23	0	1
English	Dummy = 1 for native English speaker, 0 for accented instructor	0.62	0.24	0	1
Full time	Dummy = 1 for full time, 0 for contract instructor	0.52	0.25	0	1
Male	Dummy = 1 for male, 0 for female instructor	0.78	0.17	0	1
Notes	Dummy =1 if instructor provides notes, 0 otherwise	0.53	0.25	0	1
Retired	Dummy = 1 for retired faculty, 0 otherwise	0.02	0.02	0	1
Prof	Dummy = 1 for Associate Professor and Full Professor, 0 otherwise	0.20	0.16	0	1
Instructor	Instructor number from 1 – 80	26.10	300.68	1	80
Year	Years from 2015-2019	2017.49	1.14	2015	2019

Table 4: Regression Results

Variables/Models	Random effects panel data	Random effects panel data	Random effects panel data	Random effects panel data	Random effects with logit	Random effects with logit (Dummy = 1 for instructor rated > 4)	Random coefficients multilevel model
Diff	-0.31*** (0.05)	-0.31*** (0.05)	-0.41*** (0.06)		-0.82*** (0.18)	-0.70*** (0.14)	-0.29*** (0.06)
100 level	-0.16 (0.12)	-0.16 (0.12)	-0.23 (0.12)	-0.12 (0.13)	-0.64 (0.44)	-0.39 (0.31)	-0.17 (0.11)
Attendance	0.10 (0.08)	0.10 (0.09)	0.12 (0.08)	0.04 (0.08)	-0.06 (0.36)	0.38 (0.28)	0.09 (0.08)
Text	-0.13 (0.09)	-0.13 (0.09)	-0.13 (0.09)	-0.18* (0.10)	-0.17 (0.35)	-0.22 (0.27)	-0.12 (0.09)
Grade	0.55*** (0.10)	0.55*** (0.10)		0.74*** (0.10)	1.19*** (0.29)	1.08*** (0.27)	0.48*** (0.10)
Econ	0.27 (0.31)	0.27 (0.33)	0.24 (0.33)	0.22 (0.33)	0.52 (0.84)	0.58 (0.69)	0.29 (0.33)
Uni	-0.10 (0.32)	-0.09 (0.31)	-0.03 (0.34)	-0.22 (0.31)	-0.08 (0.82)	-0.05 (0.67)	-0.10 (0.31)
White	0.58** (0.28)	0.58** (0.28)	0.65** (0.30)	0.56** (0.28)	1.79** (0.90)	0.76 (0.73)	0.57** (0.27)
English	0.25 (0.24)	0.25 (0.25)	0.30 (0.25)	0.23 (0.24)	0.70 (0.91)	0.22 (0.72)	0.23 (0.24)
Full time	0.28 (0.20)	0.29 (0.23)	0.30 (0.21)	0.16 (0.19)	1.08 (0.64)	0.30 (0.50)	0.30 (0.19)
Male	-0.38* (0.23)	-0.38* (0.23)	-0.41* (0.23)	-0.32 (0.22)	-0.63 (0.76)	-1.05* (0.58)	-0.29 (0.21)
Notes	0.02 (0.23)	0.02 (0.23)	0.04 (0.25)	0.15 (0.21)	-0.18 (0.62)	-0.15 (0.49)	-0.01 (0.22)
Retired	0.20 (0.32)	0.21 (0.37)	0.21 (0.42)	-0.04 (0.29)	1.20 (1.76)	1.63 (1.40)	0.43* (0.24)
Constant	2.47 (0.50)	2.47 (0.52)	4.70 (0.38)	0.97 (0.49)	-1.72 (1.72)	-2.00 (1.44)	3.61 (0.33)
Prof		-0.02 (0.31)					

***, **, * denote significance at the 1%, 5% and 10% levels

Notes: The standard errors are robust. Difficulty and Grade are grand mean centred for the random coefficients model.