When RateMyProfessors met Google Scholar: students' evaluations and professors' looks and research[1]

Rómulo A. Chumacero, Ricardo D. Paredes and Tomás Reyes

Received: 14/11/2022
Accepted: 01/02/2023

Abstract

We take more than a million student evaluations of almost 200,000 professors from the RateMyProfessors website and link them with information on the research productivity of almost 3,000 professors in Google Scholar to provide a systematic characterization of the relationship between student evaluations and the characteristics of the classes, universities and professors concerned and to test whether students' appreciations are conditionally related to research productivity. The study concludes that although how "easy" and "interesting" students consider a course to be are the most important determinants of their evaluations, there is a "looks" or "beauty" premium, with no systematic racial, age or gender component. Surprisingly, research productivity is either not significant or is negatively related to the assessment of a professor's teaching abilities.

Keywords

Higher education, educational quality, teachers, evaluation, students, data analysis, mathematical models, websites

JEL classification

A22, C10, C20

Authors

Rómulo A. Chumacero is an associate professor at the Department of Economics of the University of Chile. Email: rchumace@fen.uchile.cl.

Ricardo D. Paredes is a full professor at the Department of Industrial and Systems Engineering of the Pontifical Catholic University of Chile and member of the Competition Court (TDLC) of Chile. Email: rparedes@tdlc.cl.

Tomás Reyes is an associate professor at the Department of Industrial and Systems Engineering of the Pontifical Catholic University of Chile. Email: threyes@uc.cl.

[1] The authors are grateful to an anonymous referee and to Julio Pertuzé for their helpful comments and suggestions; to Valeria Acevedo, Antonia Bezanilla, Bárbara Fredes, Joaquín Lezaeta, Esteban Miño, Lorena Moraga, Patricio Niculqueo, Leticia Ortiz, Antonio Ossa, Sara Paolini, Isabella Sassi and Hernán Scheel for their valuable research assistance; and to Fondecyt projects 1171894, 1140980 and 1211367 for financial support. The usual disclaimer applies.

I. Introduction

Most higher education institutions use student evaluation of teaching (SET) instruments to assess performance on the courses they offer (Dommeyer, Baum and Hanna, 2002; Dommeyer and others, 2002). SET instruments are important because they provide teachers with immediate feedback and information on how students perceive their classes, which they can use to improve their performance. They also provide information to academic managers responsible for assigning courses and to prospective future students about a teacher's characteristics. These are some of the reasons why SET instruments have been employed by staff administrators to determine promotions, while accreditation agencies have favoured them as a proxy for good academic management.

However, although SET instruments provide a metric of student opinions, they can be misused (Jones, Gaffney-Rhys and Jones, 2014). Individual evaluations may be determined by factors different to those that experts associate with teaching quality and, more importantly, may be subject to systematic biases and lack of validity (Onwuegbuzie, Daniel and Collins, 2009). Thus, students may value a course that is relatively easy over a rigorous one, or favour a personal characteristic of the professor, such as appearance, that is not necessarily related to quality.

On the management side, given that some professors also conduct research, a proper evaluation by faculty administrators should consider the relationship between the two activities. Professors face a trade-off between time spent on research and teaching (although it may be the case that time spent on research is a valuable input for teaching). Whether research contributes to undergraduate teaching, as public policies that subsidize research from student fees may suggest, needs to be tested empirically. Furthermore, Carlozzi (2018) warns about confirmation bias in research, something that may also affect administrative decisions. When classifying researchers by their attitude towards SET instruments, he found that lead authors with negative views on these were 14 times as likely to score below a predicted estimate, based on RateMyProfessors (RMP) evaluations.

Moreover, SET instruments are usually summarized in a descriptive statistic (average), which may mask relevant information and lead to mistaken conclusions. Misuse of SET instruments can adversely affect the work environment and create grade inflation (Abrami and Mizener, 1985; Dowd, 1988; Goldman, 1990; Blunt, 1991; Marsh and Bailey, 1993; Benson and Lewis, 1994; Johnson, 2003; Onwuegbuzie, Daniel and Collins, 2009).

As SET instruments are not usually accessible to students, informal evaluations have emerged through open websites. RMP is probably the most popular, with more than 19 million evaluations of 1.7 million teachers from over 7,500 institutions (RateMyProfessors, 2020). Its value as a source of teacher ratings relies on a reasonably high correlation with institutional evaluations (Sonntag, Bassett and Snyder, 2009).

This paper uses that database, which is substantially larger than those employed by most studies, to assess the relevance of different variables in the way students evaluate their teachers and to relate their evaluations to their subjective impressions and the academic productivity of the professors concerned. The paper's main methodological contribution is to link the evaluations made in RMP with professors' academic record from Google Scholar (GS). This association allows us to evaluate the relationship between a professor's role as a teacher and as a researcher. We also analyse the written comments left by students on the RMP website to identify positive (affectionate, uplifting, etc.) or negative (resentful, angry, etc.) responses from the tone of the language used. With this information, we can evaluate the main determinants of the average evaluation obtained by a professor on RMP.

The rest of the paper is organized as follows. Section II briefly reviews the literature on RMP evaluations. Section III provides a simple theoretical model that explains the results obtained here. Section IV describes the dataset used. Section V presents the main statistical properties of the evaluations and conducts statistical exercises to better characterize the determinants of the evaluations. Lastly, section VI concludes.

II. Literature review

The debate on SET instruments gathered strength in the United States and Canada with the growing popularity of evaluations through open websites, with RMP being probably the most popular. RMP includes five variables in each evaluation: (i) "easiness", meaning the ability to get good grades without much effort; (ii) "helpfulness", meaning teachers' willingness to answer questions, particularly outside the classroom; (iii) "clarity" in classroom exposition; (iv) "overall quality"; and (v) "hotness", associated with looks or beauty.[2] The first three variables are measured on a Likert scale from 1 to 5. Overall quality is an average of helpfulness and clarity, while hotness is a dummy variable associated with the specific question of whether the student considers the teacher physically attractive (RateMyProfessors, 2020).

RMP has not been free of criticisms. Sonntag, Bassett and Snyder (2009) analysed its validity by comparing the website with evaluations carried out through the Individual Development and Educational Assessment (IDEA) survey. Using over 600 RMP evaluations for 126 teachers at Lander University, they found a Pearson correlation of 0.69 between the IDEA and RMP quality scores.[3] The easiness dimension and the average grade were also highly correlated. The authors concluded that RMP yielded results similar to those of institutional evaluations. Using a sample of 399 professors in RMP, Otto, Sanford and Ross (2008) found a positive correlation between clarity and helpfulness and a negative correlation between these variables and the variance in easiness.[4]

Boswell (2016) compared the effects of students' RMP evaluations and university-administered evaluations on teacher's self-efficacy, defined as the set of beliefs that determine how well a person can execute a plan of action in prospective situations, and concluded that while teachers thought RMP was less accurate and less serious, the feedback provided by the two evaluations influenced professors equally.

Some researchers have reported a "halo effect" in RMP evaluations, meaning that only students with extreme opinions participate (Felton and others, 2008; Clayson, 2013). The timing of evaluations also seems to be an important variable. Using a sample of 25 teachers at a medium-sized university in the United States, Legg and Wilson (2012) found that evaluations carried out at the beginning of a course yielded lower scores than those carried out at the end, and that the "easiness" results changed depending on the timing of the assessment.

[2] RMP was founded in May 1999 by John Swapceinski. It was acquired in 2005 by Patrick Nagle and William DeSantis, who resold it in 2007 to Viacom. Cheddar acquired RMP in 2018. During this time, RMP has made changes to the format of its questionnaire. We construct our database from information available up to 2018.

[3] A value of 1 in the Pearson correlation index represents a perfect positive correlation and a value of -1 a perfect negative correlation.

[4] A related literature has evaluated similarities and differences in evaluations of face-to-face versus online courses (Rovai and others, 2006; Kelly, Ponton and Rovai, 2007; Dziuban and Moskal, 2011).

Using the evaluations of 94 teachers at the University of Texas, Hamermesh and Parker (2005) found beauty to have a highly significant effect on scores. Similarly, Bonds-Raacke and Raacke (2007), Johnson and Crews (2013), Freng and Webber (2009) and Riniolo and others (2006) found a positive conditional correlation between overall class quality and beauty. Indeed, Freng and Webber (2009) suggested that "hotness" accounted for a substantial part of the variance in quality ratings.

Considering a larger sample, Felton and others (2008) replicated previous studies and found a correlation of 0.6 between hotness and overall class quality, which is over twice as high as that identified in earlier work (Felton, Mitchell and Stinson, 2004). Sen, Voia and Woolley (2010) used a sample of economics professors at 16 universities and their RMP information to evaluate the effects of "hotness" on salaries, teaching quality and research productivity. They found that "hotness" generated significant earnings premiums and was highly correlated with teaching productivity but not research productivity.

Green, Mixon and Treviño (2005) presented evidence for self-selection, with more attractive prospective professors tending to choose more liberal arts-oriented colleges or universities over research-oriented colleges or universities. Mixon and Smith (2013) took a random subset of 200 professors from the RMP data and concluded that more attractive professors traded on their appearance by offering more rigorous courses.

Regarding the relationship between research and teaching, different hypotheses exist. One stresses scarcity of resources resulting in a trade-off between the time spent on research and that used to prepare classes (see, for instance, Fox, 1992; Cretchley and others, 2014; Walstad and Allgood, 2005; Arnold, 2008). A negative correlation between research and teaching quality could also be the result of specialization, with some professors specializing in teaching and others in research in the light of comparative advantages (see, for instance, Hollywood and others, 2020).

In contrast, other authors argue that most professors consider research a valuable input for good teaching. Barnett (1992) suggest that research, in a context of discovery, is vital for good teaching (see also Neuman, 1992; Becker and Kennedy, 2005). Uz Zaman (2004) argues for complementarity between research and teaching, resulting, among other things, from the increase in critical thinking abilities that research generates. In the same vein, McCaughey (1994) reports that most academics recognize that their research agenda shapes their teaching programmes.

Ramsden and Moses (1992) find that a commitment to teaching correlates negatively with research; however, differing information sources and small samples mean that the empirical relationship between teaching and research is not conclusive. Feldman (1987) reviews 43 studies and finds a weak positive correlation between teaching as assessed by students and research productivity. Hattie and Marsh (1996), in a meta-analysis based on 58 studies, conclude that there is no correlation between research and teaching. Cadez, Dimovski and Zaman Groff (2017) find that research productivity is not related to teaching quality, but that research quality is positively related to teaching quality. Palali and others (2018) find for the Netherlands that a better research record is not reflected in master's students' evaluations and that bachelor students give lower scores to teachers with better research.

Lastly, teachers' age and gender have also proved to be important factors in teaching evaluations. Stonebraker and Stone (2015) found that age negatively affected teaching evaluations. However, when they restricted the sample to the professors students found hot, the age effect vanished. Meanwhile, using students' questionnaires, Clayson (2020) found that students learnt more from older professors but that younger ones were more helpful. She also found that students preferred male to female professors, particularly when they focused on learning.

III. A simple model

This section presents a simple model to illustrate the possible interactions between the time spent on research and teaching activities.

Consider a professor interested in maximizing his or her utility, which depends on consumption (c) and leisure (l):

$$max\, u(c,l) \tag{1}$$

subject to the constraints:

$$\begin{aligned} c &\leq y(t, r, w, x), \\ 1 &= l + t + r, \\ t &\geq t_0, \\ r &\geq r_0 \end{aligned} \tag{2}$$

where y is the income generated by teaching (t) and research (r) activities, while w and x are their respective compensation. It is also assumed that the individual has one unit of time which has to be divided between leisure, teaching and research. Lastly, t_0 and r_0 are the minimum time required by the administrator to be spent on each activity.[5]

From the first order conditions of the maximization problem, we obtain:

$$u_c y_t + k_t = u_c y_r + k_r \tag{3}$$

where v_z represents the derivative of function v ($v = u,y$) with respect to variable z ($z = c,t,r$) and k_z is the multiplier associated with the constraint for time spent on activity z ($z = t,r$).

This simple model has interesting implications for the decision as to how to allocate time between teaching and research. For example, consider that income y is determined by:

$$y(t, r, w, x) = wt + xr \tag{4}$$

Furthermore, assume that $w > x$; i.e., the compensation for teaching activities exceeds that for time spent on research. As no special value (in terms of preferences) is placed on time spent on research, the teacher devotes r_0 of his or her time (the minimum required) to research. In that case, the time spent on teaching must satisfy:

$$u_c w + k_t = k_r > 0 \tag{5}$$

If the optimal time spent on teaching (subject to the constraint that $r^* = r_0$) is such that it exceeds the minimum required ($t^* > t_0$), it can be determined by maximizing (1) subject to (4) and $r^* = r_0$. Otherwise, the constrained optimization would imply that $t^* = t_0$. Conversely, if $w < x$, then $t^* = t_0$ is the time spent on teaching, and what must be determined is whether the optimal time spent on research is r_0 or whether it exceeds this.

[5] Instead of considering t and r as time spent on each activity, these variables could reflect their respective quality.

In either case, if remuneration is greater for one activity than the other, any amount of time that exceeds the minimum required teaching and research time will be spent on the activity with higher compensation. In the event that $w = x$, since the teacher has no preference between the two activities, he or she will either devote the minimum required to both (if the time constraints are binding) or will be satisfied with any combination of the two, as in this case they are perfect substitutes.

The examples above spell out the obvious: time spent on one activity implies less time for the other. If spending more time on one activity results in professors getting better at it, the fact that there is a trade-off between the two means that the quality of their teaching and thence their evaluations may be affected as a result.

Lastly, consider the case in which time spent on research can help to improve the quality of teaching, i.e.:

$$y(t, r, w, x) = w(r)t + xr \quad (6)$$

where $w(r)$ depends on r. Thus, the productivity of teaching also depends on research.

The maximization of (1) subject to (2) and (6) leads to:

$$u_c w + k_t = u_c(w_r + x) + k_t \quad (7)$$

where w_r is the derivative of w with respect to r. If $w_r > 0$, it is possible to have internal solutions for the time spent on teaching and research, even if $w > x$. In this case, the link between time spent on research and the "quality" of teaching may ameliorate the time trade-off.

Summarizing, this section presents a simple model that sets out the linkage between time spent on teaching, its quality, and time spent on research. As is clear, there is a trade-off in terms of time spent on each activity (and presumably a quality trade-off too). There may be a positive association between time spent on research and the quality of teaching, with the experience gained by researching improving the quality of teaching and more than compensating for the time trade-off.

IV. The data

This section describes the data and sample construction. Data were collected from two main sources: the RateMyProfessors.com website and GS academic profiles.

1. The RMP database

RMP is a popular platform where students evaluate their professors. The evaluations and comments are publicly available on the website. We used a web crawler to download information on 1,281, 193 professors rated on RMP. For all professors, we obtained their name, the name and location of the university where they taught, the department to which they were affiliated, the number of student evaluations, and whether students found them attractive.

Many of these professors had very few evaluations, so we kept only those with 20 or more in order to obtain representative indicators.[6] This criterion left 197,037 professors in the sample, and we downloaded specific data from the RMP database for each of these, such as the overall quality of the professor, how easy and interesting students found their classes, and written comments.

[6] Since our analysis focuses on the average score received by each professor, 20 is a large enough number of observations to provide a reliable estimate of the central trend of the evaluations.

We then used the Linguistic Inquiry and Word Count (LIWC) software to analyse each written comment. The software provides a score for different language dimensions according to the percentage of words relating to that dimension. Dimensions include "anger" (identified from words such as "hate" and "annoyed"), "positiveness" (words such as "nice" and "sweet") and "negativeness" (words such as "hurt" and "nasty").

We classified each professor's department into standardized faculties or schools. The 930 different departments professors belonged to were manually classified into categories such as architecture and planning, engineering, humanities, arts, social sciences, and business and management.[7] Since gender information is not provided by RMP, we used an algorithm to assign a gender to each professor based on text analysis of his or her name.[8]

We supplemented the RMP data with two other sources. First, we found that the 197,037 professors identified from RMP belonged to 2,583 different universities, and we matched each of these universities to the World Ranking Web of Universities (WRWU).[9] Second, we looked for each professor on GS to see if he or she had an academic profile available, and if that was the case we proceeded as described in the next section.[10]

2. The GS database

We complemented the data from RMP with data from GS (scholar.google.com). In GS, professors can create profiles and make them available to track their citations and manage their academic articles.

We used a script to automate the process of pairing each RMP professor with a GS profile.[11] This methodology yielded 2,401 perfect matches in which both the professor's and the university's names were written identically. The methodology also threw up multiple imperfect matches, and we manually checked the 10,000 most likely ones, finding 434 additional matches.[12]

The final list of professors for whom information was available on both RMP and GS included 2,835 individuals. For each of them, we obtained a variety of information from GS such as the number of articles the professor had published, the number of times these articles had been cited, and his or her h-index and i10-index.[13]

Lastly, we supplemented these data with information gathered manually about each professor's appearance. For each of the 2,835 professors, we collected two of the most recent photos appearing on the Internet and evaluated physical attributes such as skin colour, eye colour, age and weight. While there are different approaches to assessing beauty, including related aspects such as dominance (see Paredes, Pino and Díaz, 2019), we selected 10 research assistants to share this highly time-consuming task, employing several mechanisms to keep their criteria consistent throughout the process.[14] Each

[7] We were unable to classify 18 departments (e.g., "Honours", "Graduate Studies").

[8] The algorithm is based on the dictionary from the Gender programme written by Jörg Michael. This contains more than 40,000 first names, covering most names used in the United States, Europe, China, India and Japan. We did not get a match for 7.6% of the names; in these cases, the gender was deemed to be unknown (e.g., "G.R. Williams", "Professor Agarwal", "Yartz").

[9] We found 1,925 exact matches, in which the university names reported by RMP and WRWU matched perfectly. For the remaining 658 university names, we manually searched for the best possible match. We were unable to find a match for 182 universities.

[10] The list of variables and their definitions for the sample of 197,037 professors from RMP are given in annex A1.

[11] Pairing the information from the two databases was not trivial, since professors' names could be written differently. For each professor, the script first searched GS for the name as reported by RMP. It then compared the two names and his or her university, as reported by RMP and GS, and assigned scores depending on how similar they were.

[12] These 434 manually accepted matches were not recognized properly by the script, mainly owing to the use of abbreviations in the input texts (e.g., "Sam" instead of "Samuel", "Penn State University" instead of "Pennsylvania State University").

[13] Both indices are designed to capture the quantity and quality of publications. The h-index is an author-level metric that measures both productivity and the citation impact of publications. It is based on the set of most cited papers and the number of citations they have received in other publications. The i10-index for an author measures the number of publications that have at least 10 citations.

[14] Having an exogenous assessment of a professor's looks by research assistants yields a metric for "beauty" that is not potentially contaminated by other characteristics of the professor's which may affect the evaluations of students who know him or her.

research assistant received a list of professors and a handbook with specific instructions on how to tabulate each physical attribute.[15] We provided answers to common questions research assistants might have and asked each research assistant to review 10 professors first. We then reviewed their initial work, gave them feedback and asked them to continue, leaving a comment about any attribute of a professor that caused them doubts. We then took a random sample of 150 professors to check their work. We noticed that most inconsistencies related to attributes marked with a comment and so decided to review all attributes thus marked in the full sample.[16]

V. Results

1. Results from the RMP database

The RMP database provides information that is useful for gauging how "overall teaching quality" as measured by student assessments relates to professors' personal and academic characteristics and to the universities and schools where they teach. We are particularly interested in two aspects included in the RMP database: students' assessment of "easiness" and of how "interesting" a course is.

We first consider the whole database obtained from the RMP website (RMP database). The main advantage of this database is its size, while its main drawback is that it does not have detailed information on professors' personal and academic characteristics. For example, while RMP asks students to evaluate whether a professor is "hot", it does not report other basic characteristics such as the professor's gender.

To assess professors' academic characteristics, we searched for them in Google Scholar and matched them with the information found there. Google Scholar has more precise personal information on professors featured there in addition to their academic achievements, such as the impact factor of research, papers published, etc. Thus, for instance, if a photograph of the professor is available, we use it to obtain more physical characteristics than can be gleaned from the dichotomous "hot or not" question of the RMP database.

Table 1 presents summary statistics for the variables considered to be determinants of professors' evaluations in the RMP database. In addition to the usual summary statistics, the last column presents the Pearson correlation of each variable with the average quality of professors.

As can be seen in table 1, 90.3% of the professors evaluated teach in the United States, 9.6% in Canada and most of the remainder in the United Kingdom. Professors' evaluations are negatively correlated with their university's ranking, i.e., professors at better-ranked universities have (on average) marginally worse ratings.[17] This result also holds when we consider professors at the 50 top-ranked universities. Male professors are (marginally) better evaluated and professors appearing on Google Scholar (marginally) worse, although they account for only 1.4% of the professors in the database.

Interestingly, a personal characteristic that strongly correlates with the assessment of a course's quality is the teacher's "hotness". Likewise, "easier" courses tend to be evaluated better. How interesting students perceive a course to be also appears to make an important contribution to positive evaluations, however.

[15] See annex A2 for this material.

[16] The additional variables available for the sample of 2,835 professors with information from GS and their definitions are listed in annex A3.

[17] A better-ranked institution has a lower number associated with it for the Ranking variable.

Table 1
Descriptive statistics for professors' ratings in the RateMyProfessors database

	Mean	Median	Standard deviation	Correlation
University characteristics				
United States	0.903	1.000	0.295	0.051
Canada	0.096	0.000	0.295	-0.051
Ranking	2 767.512	1 809.000	2 686.703	0.097
Top 50	0.070	0.000	0.255	-0.013
Personal characteristics				
Male	0.658	1.000	0.474	0.005
Hotness	0.077	0.000	0.175	0.402
Academic characteristics				
Google Scholar	0.014	0.000	0.119	-0.007
Ratings				
Quality (average)	3.676	3.791	0.827	1.000
Easiness (average)	3.104	3.103	0.773	0.601
Interesting (average)	3.357	3.368	0.540	0.526
Anger (average)	0.497	0.353	0.545	-0.398
Anxiety (average)	0.368	0.227	0.519	-0.421
Certainty (average)	2.159	2.084	0.717	0.119
Negative (average)	2.250	1.894	1.474	-0.669
Positive (average)	8.023	7.680	3.256	0.717
Sadness (average)	0.268	0.217	0.269	-0.240
Tentative (average)	3.047	3.026	0.827	-0.328
Word count (average)	37.059	37.911	8.319	-0.123
Words per sentence (average)	12.922	12.818	2.321	-0.271

Source: Prepared by the authors.
Note: Correlation denotes the Pearson correlation coefficient between a variable and the average overall teaching quality of the course.

The comments left by students on the RMP website were also available to us, and we found that the more negative responses were (i.e., the more they conveyed anger or anxiety, or the more hesitant they were, for example), the worse the average student assessment was. Meanwhile, positive comments (i.e., comments containing positive feedback or words conveying certainty) correlated positively with students' assessments.

However, correlations are not enough to describe the determinants of student assessments of course quality. Furthermore, since a number of these variables are correlated with one another, correlations do not serve to ascertain their relative importance.

Table 2 presents the results of a linear regression, with the average overall quality rating in student assessments as the dependent variable. This regression shows that the geographical location of the university is not a relevant characteristic in students' evaluations, but its ranking is. The relationship is non-linear, since while better-ranked universities tend to be initially associated with poorer evaluations, after a certain ranking (approximately 6,000, out of almost 12,000) the relationship reverses.

Table 2
Determinants of overall teaching quality ratings in the RateMyProfessors database

Variable	Coefficient	Standard error	t-statistic	Probability
Constant	0.8785	0.1869	4.699	0.000
United States	0.1129	0.1350	0.836	0.403
Canada	-0.0269	0.1350	-0.199	0.842
Ranking	9.69E-06	1.18E-06	8.239	0.000
Ranking squared	-8.43E-10	1.29E-10	-6.536	0.000
Top 50	-0.0191	0.0042	-4.546	0.000
Male	0.0117	0.0021	5.686	0.000
Hotness	0.4864	0.0080	60.614	0.000
Google Scholar	0.0134	0.0083	1.617	0.106
Easiness	0.8672	0.0455	19.060	0.000
Easiness squared	-0.1735	0.0148	-11.705	0.000
Easiness cubed	0.0141	0.0016	9.035	0.000
Interesting	-0.3356	0.1123	-2.989	0.003
Interesting squared	0.2971	0.0341	8.705	0.000
Interesting cubed	-0.0400	0.0034	-11.750	0.000
Anger	-0.0120	0.0028	-4.245	0.000
Anxiety	-0.0630	0.0033	-18.860	0.000
Certainty	0.0438	0.0015	29.106	0.000
Negativeness	-0.1208	0.0016	-77.031	0.000
Positiveness	0.0949	0.0006	154.459	0.000
Sadness	-0.0382	0.0049	-7.861	0.000
Tentativeness	-0.0723	0.0014	-50.469	0.000
Word count	0.0141	0.0002	78.864	0.000
Words per sentence	-0.0285	0.0006	-51.046	0.000
Architecture	-0.1058	0.0208	-5.086	0.000
Engineering	-0.0370	0.0078	-4.722	0.000
Humanities	0.0490	0.0060	8.192	0.000
Business	0.0777	0.0066	11.758	0.000
Science	0.1326	0.0062	21.236	0.000
Health	-0.0160	0.0095	-1.682	0.093
Other	-0.0465	0.0127	-3.647	0.000
Not classified	0.0681	0.0361	1.888	0.059
R-squared	0.7330	Mean dependent variable		3.676
Adjusted R-squared	0.7330	Standard deviation dependent variable		0.827
Standard error of regression	0.4270	Akaike information criterion		1.138
Residual sum of squares	35 990	Schwarz criterion		1.140
Log likelihood	-112 085	Hannan-Quinn criterion		1.139
F-statistic	17 448	Probability (F-statistic)		0.000

Source: Prepared by the authors.
Note: White heteroskedasticity-consistent standard errors. Number of observations included: 197,029.

Male professors tend to be better evaluated, but although the coefficient is significant, its magnitude is not that great (0.0117 points more on a scale from 1 to 5). Appearing in Google Scholar is not statistically significant. By contrast with these minor effects, "hotness" is an important predictor of student ratings. The difference in the average quality rating between a situation in which no student considers a professor "hot" and one in which everyone does is 0.4864 points. Whether this means that “good-looking” professors are “better” at teaching or that there is a “beauty” premium is abundantly discussed in the literature (Hamermesh, 2011). In any event, the “beauty” premium appears to be more important than the other personal characteristics considered.

As regards “easiness” and how “interesting” courses are, we also find a non-linear relationship. While “easier” courses are better rated, there is a diminishing premium on “easiness”. By contrast, there is an increasing premium for how “interesting” a course is, up to a certain point, after which it decreases.[18] More importantly, “easiness” contributes marginally more to a good evaluation than how “interesting” a course is.

Regarding feelings and emotions expressed in the written comments, as expected, negative feedback tends to be associated with a lower average quality and positive feedback with the opposite. Presumably students who write comments have more extreme opinions. For example, an average of one more word signalling anger is associated with 0.012 of a point less in the average evaluation. Interestingly, negative feedback is “numerically” more important in absolute terms. Thus, a negative comment penalizes an evaluation more than a positive comment enhances it.[19]

Lastly, table 2 also shows the faculties where professors teach. As the categories used are qualitative, the results reported take professors in faculties of education as a basis of comparison. The evidence shows that professors in business, humanities and science faculties tend to be better evaluated and those in engineering, architecture and health faculties marginally worse evaluated.

Summarizing, when all the observables are considered together, some factors are more important than others in describing what lies behind students’ perception of the quality of the courses they take. By far the most important determinants are, first, how “easy” and, second, how “interesting” a course is. Excluding “hotness”, the other factors are not statistically significant or are only weakly correlated with the average assessment. Notably, neither research as measured by a dummy variable indicating whether the professor appears in GS nor a better university ranking is an important factor in predicting how students evaluate the quality of teaching.[20]

2. Results when the RMP database is merged with the GS database

Besides the evaluation of all professors in the RMP database that has already been considered, two of the aspects mentioned above merit a more thorough examination: how important appearance (“hotness”) is and what the relationship between teaching and research is. The RMP database suggests that “hotness” is relevant and more important than, for example, the ranking of the university or whether the professor conducts research.

To analyse these aspects in depth, we shall now consider only the sample of professors appearing in GS. This reduces the number of observations to 2,835 professors out of the initial 197,037. One benefit of this reduction is that we can more precisely measure teachers’ academic productivity (as measured by the impact factors of their research). We also have access to more information on their looks from their profiles and photos. The impact factors can help provide a better understanding of the relationship between research and teaching for professors who are actively involved in the former. The photographs can provide a rough impression of “looks” uncontaminated by other characteristics of the person (such as personality, self-confidence, etc.) that influence the perception students have. They also provide more detailed knowledge of professors’ physical characteristics.

[18] Average quality assessment is maximized when the average evaluation is approximately 4.5 (out of 5).

[19] It could be that the direction of causality is the opposite, i.e., emotions expressed in describing a course may be a result of its quality.

[20] Of course, students at universities with different rankings are not directly comparable.

Table 3 shows the descriptive statistics for the variables considered, and table 4 shows the results of a regression that includes variables capturing physical attributes and academic characteristics (publications and impact factors) as determinants of average overall quality.

Table 3
Descriptive statistics derived from the Google Scholar database

	Mean	Median	Standard deviation	Correlation
University characteristics				
United States	0.799	1.000	0.401	0.061
Canada	0.201	0.000	0.401	-0.063
Ranking	854.160	370.000	1 378.762	0.081
Top 50	0.155	0.000	0.362	-0.046
Personal characteristics				
Male	0.799	1.000	0.401	0.051
Aged less than 40	0.193	0.000	0.394	0.080
Aged between 40 and 55	0.534	1.000	0.499	0.069
Bald	0.144	0.000	0.351	-0.019
Friendly	0.782	1.000	0.413	0.078
Bright eyes	0.404	0.000	0.491	0.074
Caucasian	0.854	1.000	0.353	0.124
African-American	0.018	0.000	0.132	-0.053
Indian	0.070	0.000	0.255	-0.036
Pale skin	0.621	1.000	0.485	0.050
Brown skin	0.303	0.000	0.460	-0.020
Black hair	0.251	0.000	0.433	-0.050
Blond hair	0.105	0.000	0.306	0.013
Brown hair	0.359	0.000	0.480	0.078
Grey hair	0.259	0.000	0.438	-0.064
Thin	0.061	0.000	0.239	-0.012
Average build	0.759	1.000	0.428	0.027
Tall	0.131	0.000	0.338	0.081
Average height	0.194	0.000	0.396	0.045
Good looks	0.168	0.000	0.374	0.120
Average looks	0.668	1.000	0.471	-0.012
Hotness	0.082	0.000	0.177	0.404
Academic characteristics				
Number of publications	82.481	43.000	187.503	-0.123
Citations	2 166.393	696.000	5 119.203	-0.067
H-index	16.150	13.000	13.839	-0.125
I10-index	28.378	15.000	49.647	-0.108
Evaluations				
Quality (average)	3.631	3.722	0.765	1.000
Easiness (average)	2.928	2.933	0.661	0.521
Interesting (average)	3.399	3.419	0.496	0.489
Anger (average)	0.494	0.363	0.532	-0.345
Anxiety (average)	0.385	0.233	0.559	-0.406
Certainty (average)	2.124	2.027	0.726	0.123
Negative (average)	2.428	2.096	1.514	-0.632
Positive (average)	8.118	7.671	3.294	0.668
Sadness (average)	0.267	0.213	0.274	-0.271
Tentative (average)	3.148	3.137	0.854	-0.298
Word count (average)	36.708	37.700	8.921	-0.113
Words per sentence (average)	12.635	12.602	2.249	-0.277

Source: Prepared by the authors.
Note: Correlation denotes the Pearson correlation coefficient between a variable and the average overall teaching quality of the course.

Table 4
Determinants of overall teaching quality derived from the Google Scholar database

Variable	Coefficient	Standard error	t-statistic	Probability
Constant	0.3558	0.3081	1.155	0.248
United States	0.0754	0.0854	0.883	0.377
Canada	-0.0845	0.0879	-0.961	0.337
Ranking	3.43E-05	1.63E-05	2.110	0.035
Ranking squared	-2.88E-09	1.84E-09	-1.561	0.119
Top 50	-0.0257	0.0263	-0.978	0.328
Male	0.0521	0.0235	2.217	0.027
40 or younger	0.1157	0.0307	3.768	0.000
Between 41 and 55	0.0874	0.0236	3.709	0.000
Bald	0.0301	0.0253	1.190	0.234
Brown skin	-0.0133	0.0422	-0.314	0.753
Black hair	-0.0665	0.0510	-1.302	0.193
Blond hair	-0.1462	0.0523	-2.794	0.005
Brown hair	-0.0392	0.0478	-0.821	0.412
Grey hair	-0.0731	0.0500	-1.461	0.144
Bright eyes	0.0372	0.0192	1.940	0.053
Caucasian	0.1689	0.0418	4.040	0.000
African American	-0.0380	0.0833	-0.455	0.649
Indian	0.0941	0.0526	1.789	0.074
Pale skin	0.0043	0.0434	0.099	0.921
Thin	0.0220	0.0396	0.557	0.578
Normal weight	0.0300	0.0229	1.310	0.190
Tall	0.0381	0.0265	1.435	0.151
Normal height	0.0198	0.0222	0.893	0.372
Friendly	0.0549	0.0222	2.473	0.014
Hotness	0.4449	0.0502	8.859	0.000
Good looks	-0.0102	0.0324	-0.315	0.753
Normal appearance	0.0245	0.0251	0.972	0.331
Publications	-0.0003	0.0002	-2.306	0.021
Publications squared	8.81E-08	5.54E-08	1.590	0.112
Citations	-1.70E-06	2.43E-06	-0.700	0.484
H-index	0.0011	0.0015	0.714	0.475
I10-index	-0.0001	0.0004	-0.137	0.891
Easiness	0.3124	0.0984	3.175	0.002
Easiness squared	-0.0200	0.0158	-1.262	0.207
Interesting	0.7553	0.1652	4.571	0.000
Interesting squared	-0.0689	0.0244	-2.823	0.005
Anger	0.0040	0.0218	0.183	0.855
Anxiety	-0.0711	0.0262	-2.712	0.007
Certainty	0.0535	0.0128	4.168	0.000
Negativeness	-0.1143	0.0123	-9.262	0.000
Positiveness	0.0852	0.0047	18.005	0.000
Sadness	-0.0255	0.0415	-0.615	0.538
Tentativeness	-0.0673	0.0121	-5.572	0.000
Word count	0.0118	0.0016	7.276	0.000
Words per sentence	-0.0201	0.0054	-3.727	0.000
Architecture	-0.2388	0.1077	-2.218	0.027
Engineering	-0.1230	0.0598	-2.057	0.040
Humanities	-0.0826	0.0552	-1.497	0.134
Business	0.0152	0.0581	0.261	0.794
Science	-0.0867	0.0566	-1.531	0.126
Health	-0.1119	0.0991	-1.130	0.259
Other	-0.2930	0.1758	-1.666	0.096
Not classified	0.0186	0.0949	0.196	0.845
R-squared	0.706202	Mean dependent variable		3.630257
Adjusted R-squared	0.700035	Standard deviation dependent variable		0.771407
Standard error of regression	0.422492	Akaike information criterion		1.135425
Residual sum of squares	450.7118	Schwarz criterion		1.258022
Log likelihood	-1 410.13	Hannan-Quinn criterion		1.179863
F-statistic	114.5157	Probability (F-statistic)		0.000

Source: Prepared by the authors.
Note: White heteroskedasticity-consistent standard errors. Number of observations included: 2,579.

The results are roughly consistent with those for the whole sample. There are no statistical differences in assessed quality between countries; average ratings tend to be lower in better-ranked universities; and negative emotions tend to be more correlated with worse evaluations than positive emotions with better ones. Furthermore, "easiness" has a positive but decreasing impact on the evaluation of course quality. However, one important difference is that in this sample of professors, the more "interesting" a course, the more highly valued it is (at a decreasing rate), and this is the most important contributor to the overall assessment of the course's quality. Thus, there is a difference in priorities between "easiness" and "interesting" in the case of professors who appear on GS and presumably are more active in research.

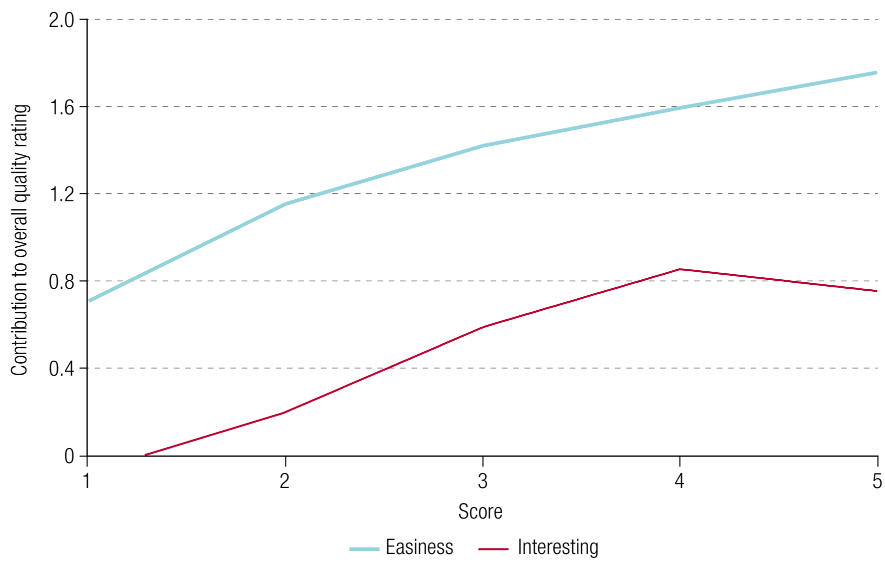
Regarding personal characteristics, male professors have slightly better ratings and "hotness" continues to be an important factor. However, very few of the specific physical characteristics mentioned in table 4 systematically help to predict the quality of a course, and in general the magnitudes are very small.

Academic productivity, whether measured by number of publications, citations or impact factors, is either irrelevant or detrimental to ratings. A possible explanation for this is that time spent on research and teaching compete.[21] Another possible explanation may be that research and teaching are complementary on graduate courses but the RMP evaluations cover mainly undergraduate courses, something the data do not allow us to differentiate. At any rate, the evidence robustly indicates that active research is not necessarily associated with better evaluations for teaching.

Figure 1 shows how the importance of factors is reversed when the RMP database is compared with the GS database. It shows, that is, that the question of how interesting the course is becomes a more important factor in the case of professors who have a research track record, as evidenced by their presence on GS.

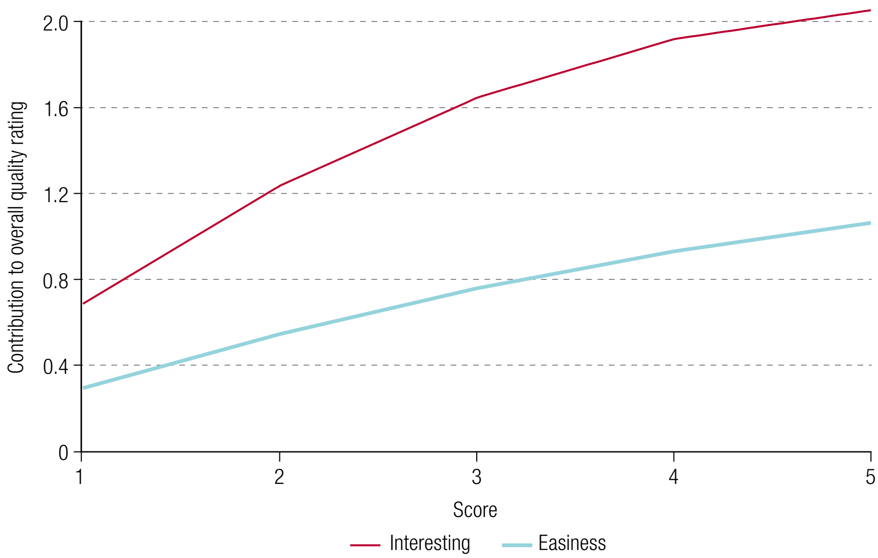
Figure 1
Contributions of "easiness" and "interesting" to professors' evaluations by students

A. RateMyProfessors database



[21] Indeed, as a referee noted, academics may be more interested in building a career based primarily on research and carry out the minimum mandatory amount of teaching.

B. Google Scholar database



Source: Prepared by the authors.
Note: The charts show the contribution made by a given score for each of the characteristics to the overall rating. For example, when the estimates of table 2 (RateMyProfessors database) are used, a score of 3 for "easiness" contributes 1.42 points to the overall quality rating, while when the table 4 estimates (GS database) are used, a score of 3 for "interesting" contributes 1.65 points to overall quality rating.

Summarizing, if we consider only professors whose research record can be tracked and ascertain further physical characteristics, we find that (qualitatively) the main results hold good. How "easy" and "interesting" courses are prove to be the main determinants of student quality ratings. There is a significant "beauty" premium that does not depend on systematic characteristics such as race or skin, eye or hair colour. It is more detrimental to professors' evaluations if they elicit negative emotions than it is beneficial if they elicit positive emotions. Lastly, there is if anything a negative association between academic productivity and teaching quality as rated by student evaluations.

VI. Concluding remarks

Whilst the use of open voluntary websites, and RMP in particular, has been criticized, there are at least three elements that make them worth describing and analysing. First, overall RMP ratings have been reported to correlate reasonably well with formal evaluations. Second, the popularity of RMP is growing, and it affects students' decisions as well as being a source of immediate feedback for professors. Lastly, it provides information for analysis of a volume and richness not available elsewhere.

In this paper, we use the RMP data set to describe the main characteristics affecting the quality of teaching, as reported by students. From the evaluations of almost 200,000 professors on the site, almost 3,000 of which could be linked to information in GS, we find that the main determinants of the perceived quality of a course are how "easy" and how "interesting" it is. How "easy" and how "interesting" a course is positively affect its quality ratings at a decreasing and increasing rate, respectively. We also find a "beauty" premium that is not systematic with respect to race, skin colour or other personal characteristics of the professors. Lastly, we find that research productivity is either insignificant or detrimental in evaluations of teaching.

Our study concludes that research seems to compete with teaching quality. If professors mainly have incentives to research, the quality of teaching may suffer. Furthermore, since students value “easiness” in their ratings, they may not be receiving the quality of education required to form competent future professionals. Exploring and understanding the reasons behind these findings may yield important implications for academic policies and decisions. Thus, the existence of a beauty premium and the possibility that an emphasis on research reduces teacher ratings give cause to reflect on the value schools set on student assessments and the costs of focusing exclusively on research.

Bibliography

Abrami, P. and D. Mizener (1985), “Student/instructor attitude similarity, student ratings, and course performance”, *Journal of Educational Psychology*, vol. 77.

Arnold, I. (2008), “Course level and the relationship between research productivity and teaching effectiveness”, *Journal of Economic Education*, vol. 39, No. 4.

Barnett, R. (1992), “Linking teaching and research: a critical inquiry”, *Journal of Higher Education*, vol. 63.

Becker, W. and P. Kennedy (2005), “Does teaching enhance research in economics?”, *American Economic Review*, vol. 95, No. 2.

Benson, D. and J. Lewis (1994), “Students’ evaluation of teaching and accountability: implications from the Boyer and the ASA reports”, *Teaching Sociology*, vol. 22.

Blunt, A. (1991), “The effects of anonymity and manipulated grades on student ratings of instructors”, *Community College Review*, vol. 18.

Bonds-Raacke, J. and J. Raacke (2007), “The relationship between physical attractiveness of professors and students’ ratings of professor quality”, *Journal of Psychiatry, Psychology and Mental Health*, vol. 1, No. 2.

Boswell, S. (2016), “Ratemyprofessors is hogwash (but I care): Effects of Ratemyprofessors and university-administered teaching evaluations on professors”, *Computers in Human Behavior*, vol. 56.

Cadez, S., V. Dimovski and M. Zaman Groff (2017), “Research, teaching and performance evaluation in academia: the salience of quality”, *Studies in Higher Education,* vol. 42, No. 8.

Carlozzi, C. (2018), “Rate my attitude: research agendas and RateMyProfessor scores”, *Assessment & Evaluation in Higher Education*, vol. 43, No. 3.

Clayson, D. (2020), “Student perception of instructors: the effect of age, gender and political leaning”, *Assessment & Evaluation in Higher Education*, vol. 45, No. 4.

___(2013), “What does ratemyprofessors.com actually rate?”, *Assessment & Evaluation in Higher Education*, vol. 39, No. 68.

Cretchley, P. and others (2014), “Research and/or learning and teaching: A study of Australian professors’ priorities, beliefs and behaviours”, *Higher Education Research and Development*, vol. 33, No. 4.

Dommeyer, C., P. Baum and R. Hanna (2002), “College students’ attitudes towards methods of collecting teaching evaluations: in-class vs. online”, *Journal of Education for Business*, vol. 78, No. 1.

Dommeyer, C. and others (2002), “Attitudes of business faculty toward two methods of collecting teaching evaluations: paper vs. online”, *Assessment & Evaluation in Higher Education*, vol. 27, No. 5.

Dowd, J. (1988), “Sociology is different: the misevaluation of teaching effectiveness”, *Sociological Inquiry*, vol. 58.

Dziuban, C. and P. Moskal (2011), “A course is a course is a course: factor invariance in student evaluation of online, blended and face-to-face learning environments”, *Internet and Higher Education*, vol. 14.

Feldman, K. (1987), “Research productivity and scholarly accomplishment of College teachers as related to their instructional effectiveness: a review and exploration”, *Research in Higher Education*, vol. 26.

Felton, J., J. Mitchell and M. Stinson (2004), “Web-based student evaluations of professors: the relations between perceived quality, easiness and sexiness”, *Assessment & Evaluation in Higher Education*, vol. 29, No. 1.

Felton, K. and others (2008), “Attractiveness, easiness and other issues: student evaluations of professors on RateMyProfessors.com”, *Assessment & Evaluation in Higher Education*, vol. 33.

Fox, M. (1992), “Research, teaching and publication productivity: mutuality versus competition in academia”, *Sociology of Education*, vol. 65.

Freng, S. and D. Webber (2009), “Turning up the heat on online teaching evaluations: does ‘hotness’ matter?”, *Teaching of Psychology*, vol. 36, No. 3.

Goldman, L. (1990), "Student evaluations of their professors rarely provide a fair measure of teaching ability", *Chronicle of Higher Education*, vol. 8, August.

Green, T., F. Mixon and L. Treviño (2005), "Have you seen the new Econ prof? beauty, teaching, and occupational choice", *Shaping the Learning Curve: Essays on Economic Education*, F. Mixon (ed.), iUniverse.

Hamermesh, D. (2011), *Beauty Pays: Why Attractive People Are More Successful*, Princeton University Press.

Hamermesh, D. and A. Parker (2005), "Beauty in the Classroom: Instructors' Pulchritude and Putative Pedagogical Productivity", *Economics of Education Review*, vol. 24.

Hattie, J. and H. W. Marsh (1996), "The relationship between research and teaching: a meta-analysis", *Review of Educational Research*, vol. 66, No. 4.

Hollywood, A. and others (2020), "Overwhelmed at first: the experience of career development in early career academics", *Journal of Further and Higher Education*, vol. 44, No. 7.

Johnson, V. (2003), *Grade Inflation: A Crisis in College Education*, New York, Springer.

Johnson, R. and A. Crews (2013), "My professor is hot! Correlates of RateMyProfessors.com ratings for criminal justice and criminology faculty members", *American Journal of Criminal Justice*, vol. 38, No. 4.

Jones, J., R. Gaffney-Rhys and E. Jones (2014), "Handle with care! An exploration of the potential risks associated with the publication and summative usage of student evaluation of teaching (SET) results", *Journal of Further and Higher Education*, vol. 38, No.1.

Kelly, H., M. Ponton and A. Rovai (2007), "A comparison of student evaluations of teaching between online and face-to-face courses", *Internet and Higher Education*, vol. 10.

Legg, A. and J. Wilson (2012), "RateMyProfessors.com offers biased evaluations", *Assessment & Evaluation in Higher Education*, vol. 37, No. 1.

Marsh, H. and M. Bailey (1993), "Multidimensional students' evaluations of teaching effectiveness: a profile analysis", *Journal of Higher Education*, vol. 64, No. 1.

McCaughey, R. (1994), *Scholars and Teachers: The Faculties of Select Liberal Arts Colleges and Their Place in American Higher Learning*, New York, Conceptual Litho Reproductions.

Mixon, F. and K. Smith (2013), "Instructor attractiveness and academic rigour: examination of student evaluation data", *Australasian Journal of Economics Education*, vol. 10, No. 1.

Neumann, R. (1992), "Perceptions in the teaching-research nexus: A framework for analysis", *Higher Education*, vol. 23.

Onwuegbuzie, A., L. Daniel and K. Collins (2009), "A meta-validation model for assessing the score-validity of student teaching evaluations", *Quality and Quantity*, vol. 43.

Otto, J., D. Sanford and D. Ross (2008), "Does RateMyProfessors.com really rate my professor?", *Assessment & Evaluation in Higher Education*, vol. 33, No. 4.

Palali, A. and others (2018), "Are good researchers also good teachers? The relationship between research quality and teaching quality", *Economics of Education Review*, vol. 64.

Paredes, V., F. Pino and D. Díaz (2019), "Does facial structure explain differences in students' evaluations of teaching? The role of perceived dominance", *Documentos de Trabajo* series, No. 483 [online] https://repositorio.uchile.cl/handle/2250/168496.

Ramsden, P. and I. Moses (1992), "Association between research and teaching in Australian higher education", *Higher Education*, vol. 23.

Rate My Professors (2020), *About RateMyProfessors.com* [online] http://www.ratemyprofessors.com/About.jsp.

Riniolo, T. and others (2006), "Hot or not: do professors perceived as physically attractive receive higher student evaluations?", *The Journal of General Psychology*, vol. 133, No. 1.

Rovai, A. and others (2006), "Student evaluation of teaching in the virtual and traditional classrooms: A comparative analysis", *Internet and Higher Education*, vol. 9.

Sen, A., M. Voia and F. Woolley (2010), "The effect of hotness on pay and productivity", *Carleton Economic Papers*, No. CEP 10-07.

Sonntag, M., J. Bassett and T. Snyder (2009), "An empirical test of the validity of student evaluations of teaching made on RateMyProfessors.com", *Assessment & Evaluation in Higher Education*, vol. 34, No. 5.

Stonebraker, R. and G. Stone (2015), "Too old to teach? The effect of age on college and university professors", *Research in Higher Education*, vol. 56, No. 8.

Uz Zaman, M. (2004), "Review of the academic evidence on the relationship between teaching and research in higher education", *Research Report*, No. RR506, London, Department for Education and Skills.

Walstad, W. and S. Allgood (2005), "Views of teaching and research in economics and other disciplines", *American Economic Review*, vol. 95, No. 2.

Annex A1

Full list of variables for the sample of 197,037 professors from RateMyProfessors

University characteristics:

- **United States:** Dummy variable that takes the value 1 if the professor's university is in the United States. Source: RMP.
- **Canada:** Dummy variable that takes the value 1 if the professor's university is in Canada. Source: RMP.
- **Ranking:** Ranking of the professor's university, with a lower number being better. Source: World Ranking Web of Universities (WRWU).
- **Top 50:** Dummy variable that takes the value 1 if the Ranking variable is less than or equal to 50. Source: WRWU.

Personal characteristics:

- **Male:** Dummy variable that takes the value 1 if the professor is male. Source: algorithm based on the gender/nam_dict dictionary written by Jörg Michael.
- **Hotness:** Dummy variable that takes the value 1 if the number of students on RMP who find the professor "hot" minus the number who do not is 20 or larger. Source: RMP.

Academic characteristics:

- **Google Scholar:** Dummy variable that takes the value 1 if the professor has a Google Scholar academic profile. Source: Google Scholar.

Evaluations:

- **Quality:** This measures a professor's quality as perceived by students. It is computed by RMP as the average between "helpfulness" and "clarity". Helpfulness is measured on a Likert scale from 1 ("No help here") to 5 ("Saved my semester"). Clarity is measured on a Likert scale from 1 ("Say what??") to 5 ("Crystal clear"). Source: RMP.
- **Easiness:** This measures how easy students find the class on a Likert scale from 1 ("Hardest thing I've ever done") to 5 ("Show up & pass"). Source: RMP.
- **Interesting:** This measures how interesting students find the professor on a Likert scale from 1 ("Meh") to 5 ("It's my life"). Source: RMP.
- **Anger:** Percentage of words in the student's comment that denote anger (words such as "hate" and "annoyed"). Source: RMP and Linguistic Inquiry and Word Count (LIWC) software.
- **Anxiety:** Percentage of words in the student's comment that denote anxiety (words such as "worried" and "fearful"). Source: RMP and LIWC.
- **Certainty:** Percentage of words in the student's comment that denote certainty (words such as "always" and "never"). Source: RMP and LIWC.

- **Negativeness:** Percentage of words in the student's comment that denote negative emotions (words such as "hurt" and "nasty"). Source: RMP and LIWC.
- **Positiveness:** Percentage of words in the student's comment that denote positive emotions (words such as "nice" and "sweet"). Source: RMP and LIWC.
- **Sadness:** Percentage of words in the student's comment that denote sadness (words such as "crying" and "sad"). Source: RMP and LIWC.
- **Tentative:** Percentage of words in the student's comment that denote tentativeness (words such as "guess" and "perhaps"). Source: RMP and LIWC.
- **Word count:** Number of words in the student's comment. Source: RMP and LIWC.
- **Words per sentence:** Average number of words per sentence in the student's comment. Source: RMP and LIWC.

Department characteristics:

- **Architecture:** Dummy variable that takes the value 1 if the professor's department is part of a school of architecture and planning. Source: manually classified on the basis of the department's name as reported by RMP.
- **Engineering:** Dummy variable that takes the value 1 if the professor's department is part of a school of engineering. Source: manually classified on the basis of the department's name as reported by RMP.
- **Humanities:** Dummy variable that takes the value 1 if the professor's department is part of a school of humanities, arts and social sciences. Source: manually classified on the basis of the department's name as reported by RMP.
- **Business:** Dummy variable that takes the value 1 if the professor's department is part of a school of business and management. Source: manually classified on the basis of the department's name as reported by RMP.
- **Science:** Dummy variable that takes the value 1 if the professor's department is part of a school of science. Source: manually classified on the basis of the department's name as reported by RMP.
- **Health:** Dummy variable that takes the value 1 if the professor's department is part of a school of health and welfare. Source: manually classified on the basis of the department's name as reported by RMP.
- **Other:** Dummy variable that takes the value 1 if the professor's department is part of a school of education, law, agriculture, counselling or sports. Source: manually classified on the basis of the department's name as reported by RMP.
- **Not classified:** Dummy variable that takes the value 1 if the professor's department could not be classified under any of the previously mentioned faculties or schools. Source: manually classified on the basis of the department's name as reported by RMP.

Annex A2

Instructions for research assistants participating in the project

It is very important that you read this entire set of instructions before starting to gather the information required.

Introduction

Each research assistant cooperating in this project will be assigned a group of professors. For each professor, you will have identifying information such as the name of the professor and the university where he/she teaches, and a unique numerical identifier assigned to that professor (ID). Based on the identifying information, you have to find two current photos of the professor and save them with names ID_1 and ID_2. Then, you have to complete a database with variables related to the physical characteristics of the professor. Below is a list of the variables; when a qualitative assessment of a professor's physical appearance is required, please be as objective as possible.

Variables

A. Gender

0. Female
1. Male

B. Skin colour

0. Pale
1. Brown
2. Dark

C. Race

0. Caucasian/White/Latin
1. African/Caribbean
2. Indian/Arab
3. Chinese/Japanese/Asian

D. Age

0. Less than 40
1. Between 40 and 55
2. More than 55

E. Eye colour

0. Black/brown
1. Green/blue

F. Hair colour

0. Black
1. Blond
2. Brown
3. Grey
4. Red

G. Bald

0. Yes
1. No

H. Weight

0. Thin/underweight
1. Average build
2. Overweight

I. Height

0. Tall
1. Average
2. Short
3. Not available

J. Looks

0. Attractive
1. Average
2. Unattractive

K. Friendly

0. Friendly
1. Unfriendly

Annex A3

Variables for the sample of 2,835 professors with a Google Scholar profile

Additional academic characteristics:

- **Number of publications:** Number of articles written by the professor. Source: Google Scholar.
- **Citations:** Number of citations of all articles written by the professor. Source: Google Scholar.
- **H-index:** Index measuring a professor's productivity and impact. A professor with an index of H has published H papers, each of which has been cited at least H times. Source: Google Scholar.
- **I10-index:** Number of articles with at least 10 citations. Source: Google Scholar.

Additional personal characteristics:

- **Less than 40:** Dummy variable that takes the value 1 if the professor is less than 40 years old. Source: visual inspection of two photographs of the professor.
- **Between 40 and 55:** Dummy variable that takes the value 1 if the professor is between 40 and 55 years old. Source: visual inspection of two photographs of the professor.
- **Bald:** Dummy variable that takes the value 1 if the professor is bald. Source: visual inspection of two photographs of the professor.
- **Friendly:** Dummy variable that takes the value 1 if the professor looks friendly. Source: visual inspection of two photographs of the professor.
- **Bright eyes:** Dummy variable that takes the value 1 if the professor has clear (e.g., green or blue) eyes. Source: visual inspection of two photographs of the professor.
- **Caucasian:** Dummy variable that takes the value 1 if the professor looks Caucasian, White or Latin. Source: visual inspection of two photographs of the professor.
- **African-American:** Dummy variable that takes the value 1 if the professor looks African-American or Caribbean. Source: visual inspection of two photographs of the professor.
- **Indian:** Dummy variable that takes the value 1 if the professor looks Indian or Arab. Source: visual inspection of two photographs of the professor.
- **Pale skin:** Dummy variable that takes the value 1 if the professor has pale or light skin. Source: visual inspection of two photographs of the professor.
- **Brown skin:** Dummy variable that takes the value 1 if the professor has brown or dark skin. Source: visual inspection of two photographs of the professor.
- **Black hair:** Dummy variable that takes the value 1 if the professor has black hair. Source: visual inspection of two photographs of the professor.
- **Blond hair:** Dummy variable that takes the value 1 if the professor has blond hair. Source: visual inspection of two photographs of the professor.
- **Brown hair:** Dummy variable that takes the value 1 if the professor has brown or brunette hair. Source: visual inspection of two photographs of the professor.

- **Grey hair:** Dummy variable that takes the value 1 if the professor has grey hair. Source: visual inspection of two photographs of the professor.
- **Thin:** Dummy variable that takes the value 1 if the professor looks thin or underweight. Source: visual inspection of two photographs of the professor.
- **Average build:** Dummy variable that takes the value 1 if the professor seems to be of average build or healthy weight. Source: visual inspection of two photographs of the professor.
- **Tall:** Dummy variable that takes the value 1 if the professor looks tall. Source: visual inspection of two photographs of the professor.
- **Average height:** Dummy variable that takes the value 1 if the professor seems to be of average height. Source: visual inspection of two photographs of the professor.
- **Good looks:** Dummy variable that takes the value 1 if the professor looks attractive. Source: visual inspection of two photographs of the professor.
- **Average looks:** Dummy variable that takes the value 1 if the professor has average looks. Source: visual inspection of two photographs of the professor.