Available online at www.sciencedirect.com
SCIENCE DIRECT®
Economics of Education Review 24 (2005) 369–376
Economics of Education Review
www.elsevier.com/locate/econedurev

Beauty in the classroom: instructors' pulchritude and putative pedagogical productivity

Daniel S. Hamermesh*, Amy Parker

Department of Economics, University of Texas, Austin, TX 78712-1173, USA

Received 14 June 2004; accepted 21 July 2004

Abstract

Adjusted for many other determinants, beauty affects earnings; but does it lead directly to the differences in productivity that we believe generate earnings differences? We take a large sample of student instructional ratings for a group of university teachers and acquire six independent measures of their beauty, and a number of other descriptors of them and their classes. Instructors who are viewed as better looking receive higher instructional ratings, with the impact of a move from the 10th to the 90th percentile of beauty being substantial. This impact exists within university departments and even within particular courses, and is larger for male than for female instructors. Disentangling whether this outcome represents productivity or discrimination is, as with the issue generally, probably impossible.

JEL classifications: J71; I29

Keywords: Beauty; Discrimination; Class evaluations; College teaching

> It was God who made me so beautiful. If I weren't, then I'd be a teacher.
>
> [Supermodel Linda Evangelista]

1. Introduction

An immense literature in social psychology (summarized by Hatfield & Sprecher, 1986) has examined the impact of human beauty on a variety of non-economic outcomes. Recently economists have considered how beauty affects labor market outcomes, particularly earnings, and have attempted to infer the sources of its effects from the behavior of different economic agents (Hamermesh & Biddle, 1994; Biddle & Hamermesh, 1998). The impacts on these monetary outcomes are implicitly the end results of the effects of beauty on productivity; but there seems to be no direct evidence of the impacts of beauty on productivity in a context in which we can be fairly sure that productivity generates economic rewards.

A substantial amount of research has indicated that academic administrators pay attention to teaching quality in setting salaries (Becker & Watts, 1999). A number of studies (e.g., Katz, 1973; Siegfried & White, 1973; Kaun, 1984; Moore, Newman, & Turnbull, 1998) have demonstrated that teaching quality generates ceteris paribus increases in salary (but see DeLorme, Hill, & Wood, 1979). The question is what generates the measured productivity for which the economic rewards

*Corresponding author. Tel.: +1 512 475 8526; fax: +1 512 471 3510.
E-mail address: hamermes@eco.utexas.edu (D.S. Hamermesh).

0272-7757/$ - see front matter © 2004 Elsevier Ltd. All rights reserved.
doi:10.1016/j.econedurev.2004.07.013

are being offered. One possibility is simply that ascriptive characteristics, such as beauty, trigger positive responses by students and lead them to evaluate some teachers more favorably, so that their beauty earns them higher economic returns.

In this study we examine the productivity effects of beauty in the context of undergraduate education.[1] In particular, we consider the impact of instructors' looks on their instructional ratings in the courses that they teach. In Section 2 we describe a data set that we have created to analyze the impact of beauty on this indicator of instructors' productivity. In Section 3 we discuss and interpret the results of studying these impacts. Section 4 presents the implications of the analysis for interpreting the impact of an ascriptive characteristic on economic outcomes as stemming from productivity effects or discrimination.

2. Measuring teaching productivity and its determinants

The University of Texas at Austin, like most other institutions of higher learning in the United States and increasingly elsewhere too, requires its faculty to be evaluated by their students in every class. Evaluations are carried out at some point in the last 3 weeks of the 15-week semester. A student administers the evaluation instrument while the instructor is absent from the classroom. The rating forms include: "Overall, this instructor was very unsatisfactory (1); unsatisfactory (2); satisfactory (3); very good (4); excellent (5);" and "Overall, this course was very unsatisfactory, unsatisfactory …." In the analysis we concentrate on responses to the second question, both because it seems more germane to inferring the instructor's educational productivity, and because, in any event, the results for the two questions are very highly positively correlated ($r = 0.95$).

We chose instructors at all levels of the academic hierarchy, obtaining instructional staffs from a number of departments that had posted all faculty members' pictures on their departmental websites. An additional ten faculty members' pictures were obtained from miscellaneous departments around the University. The average evaluation score for each undergraduate class that the faculty member taught during the academic years 2000–2002 is included. This sample selection criterion resulted in 463 classes, with the number of classes taught by the sample members ranging from 1 to 13. The classes ranged in size from 8 to 581 students (enrolled as of the 12th day of the semester, after which it becomes costly to drop a class or even switch sections in a multi-section course), while the number of students completing the instructional ratings ranged from 5 to 380. Underlying the 463 sample observations are 16,957 completed evaluations from 25,547 registered students. Both lower- and upper-division courses are included. We make this distinction because there is no way of knowing the fraction of students in a particular course for whom it is required, which might otherwise be more interesting.

We also obtained information on each faculty member's sex, whether on the tenure track or not, minority status and whether he/she received an undergraduate education in an English-speaking country.[2] Table 1 presents the statistics describing these variables and the information about the classes. The means are weighted (by the number of evaluation forms returned) averages of the individual class averages. These descriptive statistics are generally unsurprising: (1) the average class rating is below that for the instructor him/herself; (2) the average rating is around 4.0 (on the 5 to 1 scale), with a standard deviation of about 0.5; and (3) non-tenure track faculty are disproportionately assigned to lower-division courses.

Each of the instructors' pictures was rated by each of six undergraduate students: three women and three men, with one of each gender being a lower division, two upper-division students (to accord with the distribution of classes across the two levels). The raters were told to use a 10 (highest) to 1 rating scale, to concentrate on the physiognomy of the instructor in the picture, to make their ratings independent of age, and to keep 5 in mind as an average. In the analyses we unit normalized each rating. To reduce measurement error the six normalized ratings were summed to create a composite standardized beauty rating for each instructor.

Table 2 presents statistics describing the ratings of the instructors' beauty by each of the six undergraduates who did the ratings. The students clearly had some difficulty holding to the instruction that they strive for an average rating of 5, as the averages of three of the six raw ratings were significantly below that, and none was significantly above (perhaps reflecting the students' inability to judge these older people, perhaps reflecting the choices implied in the epigraph). Moreover, the standardized ratings show that five of the six sets of

[1] Linking instructors' looks to their pedagogical productivity does not appear to have been done previously, but Goebel & Cashen (1979) and Buck & Tiene (1989) did ask students in various grades to comment on the teaching ability that they would expect from individuals of varying levels of beauty based on a set of photographs. Ambady & Rosenthal (1993), the only study to look at actual teaching evaluations (of 13 TAs in a single course), focused on their non-verbal behavior but did touch on the effects of their attractiveness.

[2] This last variable is designed to account for the possibility of lower productivity of foreign teachers (see Borjas, 2000, but also Fleisher, Hashimoto, & Weinberg, 2002) that might also be correlated with perceptions of their looks. In fact, in our sample this correlation is only −0.02.

Table 1
Descriptive statistics, courses, instructors and evaluations

Variable	All	Lower division	Upper division
Course evaluation	4.022 (0.525)	4.060 (0.563)	3.993 (0.493)
Instructor evaluation	4.217 (0.540)	4.243 (0.609)	4.196 (0.481)
Number of students	55.18 (75.07)	76.50 (109.29)	44.24 (45.54)
Percent evaluating	74.43	73.52	74.89
Female	0.359	0.300	0.405
Minority	0.099	0.110	0.090
Non-native English	0.037	0.007	0.060
Tenure track	0.851	0.828	0.869
Lower division	0.339	—	—
One credit	0.029	—	—
Number of courses	463	157	306
Number of faculty	94	42	79

Note: Means with standard deviations in parentheses. All statistics except for those describing the number of students, the percent evaluating the instructor and the lower–upper division distinction are weighted by the number of students completing the course evaluation forms.

Table 2
Beauty evaluations, individual and composite

	Average	Standard deviation	Standardized	
			Minimum	Maximum
Individual ratings:				
Male, upper division—1	4.43	2.18	−1.57	2.10
Male, upper division—2	4.87	1.65	−2.34	2.50
Female, upper division—1	5.18	2.05	−2.03	1.84
Female, upper division—2	5.39	2.10	−2.10	2.20
Male, lower division	3.53	1.70	−1.49	2.04
Female, lower division	4.14	1.88	−1.67	2.05
Composite standardized rating				
	0	0.83	−1.54	1.88

ratings were skewed to the right. There was some concern, based on observations in earlier research, that the distribution of ratings of female faculty might have higher variance than that of males. While the variance was slightly higher, the Kolmogorov–Smirnov statistic testing equality of the two distributions had a p-value of 0.077.

Despite these minor difficulties, a central concern—that the assessments of beauty be consistent across raters—was achieved remarkably well. The 15 pairwise correlation coefficients of the standardized beauty ratings range from 0.54 to 0.72, with an average correlation coefficient of 0.62. Cronbach's alpha, the standard psychometric measure of concordance, is 0.91. These indicate substantial agreement among the raters about the looks of the 94 faculty members. Any disagreement or greater subjectivity about the ratings would, however, merely impart a downward bias to estimates of the impact of beauty on teaching evaluations.

3. Impact of beauty on teaching ratings

3.1. Basic results

The basic model specifies a faculty member's teaching ratings as determined by a vector of his/her characteristics, X, and by a vector of the course's characteristics, Z. Included in X are whether the instructor is female, whether he/she is a minority, whether not a native English speaker, and whether on the tenure track. The central variable in X is our composite measure of standardized beauty. Z includes whether the observation is on an upper- or lower-division course, and whether it is for one credit. (27 of the classes were one-credit labs,

Table 3
Weighted least-squares estimates of the determinants of class ratings

Variable	All	Males	Females	Lower division	Upper division
Composite standardized beauty	0.275 (0.059)	0.384 (0.076)	0.128 (0.064)	0.359 (0.092)	0.166 (0.061)
Female	−0.239 (0.085)	—	—	−0.345 (0.133)	−0.093 (0.104)
Minority	−0.249 (0.112)	0.060 (0.101)	−0.260 (0.139)	−0.288 (0.156)	−0.231 (0.107)
Non-native English	−0.253 (0.134)	−0.427 (0.143)	−0.262 (0.151)	−0.374 (0.141)	−0.286 (0.131)
Tenure track	−0.136 (0.094)	−0.056 (0.089)	−0.041 (0.133)	−0.187 (0.141)	0.005 (0.119)
Lower division	−0.046 (0.111)	0.005 (0.129)	−0.228 (0.164)	—	—
One-credit course	0.687 (0.166)	0.768 (0.119)	0.517 (0.232)	0.792 (0.101)	—
R^2	.279	.359	.162	.510	.126
N courses	463	268	195	157	306
N faculty	94	54	40	42	79

Note: Robust standard errors in parentheses here and in Table 4.

physical education or other low-intensity activities that students tend to view differently from other classes).[3] Where sample sizes permit we examine the determinants of course evaluations in lower- and upper-division courses separately, since the students in the former may be more focused on the instructor him/herself and less on the degree to which the instructor can exposit the course material.

Table 3 presents weighted least-squares estimates of the equations describing the average course evaluations. As weights we use the number of students completing the evaluation forms in each class, because the error variances in the average teaching ratings are larger the fewer students completing the instructional evaluations. We present robust standard errors that account for the clustering of the observations (because we observe multiple classes for the overwhelming majority of instructors) for each of the parameter estimates.[4]

The striking fact from the estimates in the first column is the statistical significance of the composite standardized beauty measure. The effects of differences in beauty on the average course rating are not small: Moving from one standard deviation below the mean to one standard deviation above leads to an increase in the average class rating of 0.46, close to a one-standard deviation increase in the average class rating.[5] A complete picture of the importance of beauty in affecting instructors' class evaluations is presented in Fig. 1. For instructors at each percentile of the distribution of beauty, the figure shows the class evaluation that he/she would obtain with other characteristics in X and Z at the sample means. The instructional rating varies by nearly two standard deviations between the worst- and best-looking instructors in the sample.[6]

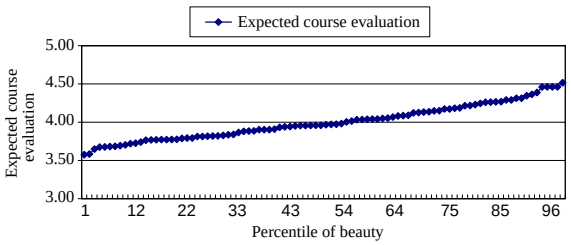

Fig. 1. Beauty and course evaluations.

[3]Age and a quadratic in age were included in other versions of the basic equation. These terms were never significantly non-zero as a pair or individually and had essentially no impact on the coefficients of the other terms in X and Z. Also unimportant was an indicator of whether the faculty member was tenured. If one-credit classes are excluded, the beauty coefficient changes slightly, rising to 0. 283.

[4]The unweighted least-squares parameter estimates differ little from those presented here. Had we failed, however, to use the correct robust standard errors, the parameter estimates here would all appear more highly significant statistically.

[5]This impact is at the intensive margin—among students who showed up in class on the day the course evaluations were completed. If we examine the extensive margin—the impact on the fraction of students attending class on that day—we also find a positive and nearly statistically significant effect of composite standardized beauty.

[6]One might be concerned about the upper and lower limits on the evaluation scores. While the lowest class average was 2.1, eight of the 463 classes did receive an average evaluation of 5.0. To examine whether this ceiling effect matters, we reestimated the basic equation in column 1 of Table 3 using an upper-limit tobit estimator. Not surprisingly, given the small fraction of observations at the ceiling, the parameter estimates were essentially unchanged. Least squares might also be problematic given the distribution of this measure. We thus also reestimated the basic equation using least absolute deviations. Again the coefficients were essentially unchanged (with the parameter estimate on the beauty measure rising slightly, to 0.299).

That inferring the impact of instructors' looks on measures of their instructional productivity requires evaluations of their looks by several raters is demonstrated by sequential reestimates of the basic equation that include each of the six raters' evaluations individually. While the class ratings are significantly related to each rater's views of the instructors, the estimated impacts range only from 0.12 to 0.23, i.e., below the estimates based on the composite standardized measure. There is substantial measurement error in the individual beauty ratings. The errors become less important once any pair of ratings is averaged: the estimated coefficients using the 15 possible pairs range from 0.19 to 0.26, and they range upward from 0.23 when any three ratings are averaged.

Minority faculty members receive lower teaching evaluations than do majority instructors, and non-native English speakers receive substantially lower ratings than do natives. Lower-division courses are rated slightly lower than upper-division courses. Non-tenure-track instructors receive course ratings that are surprisingly almost significantly higher than those of tenure-track faculty. This may arise because they are chiefly people who specialize in teaching rather than combining teaching and research, or perhaps from the incentives (in terms of reappointment and salary) that they face to please their students. The one-credit courses, all of which are lower-division, receive much higher evaluations than others, perhaps because of the nature of the courses as labs or electives.

Perhaps the most interesting result among the other variables in the vectors X and Z is the significantly lower rating received by female instructors, an effect that implies reductions in average class ratings of nearly one-half standard deviation. This disparity departs from the consensus in the literature that there is no relationship between instructor's gender and instructional ratings (Feldman, 1993).

To explore this sex difference further we estimate the basic model separately for classes taught by male and female instructors. The results are shown in columns 2 and 3 of Table 3. At the means of the variables the predicted instructional rating is lower for female instructors—the negative coefficient on the indicator in column 1 is not an artifact of a correlation of perceived beauty and gender. The reestimates show, however, that the impact of beauty on instructors' course ratings is much lower for female than for male faculty. Good looks generate more of a premium, bad looks more of a penalty for male instructors, just as was demonstrated (Hamermesh & Biddle, 1994) for the effects of beauty in wage determination.

Columns 4 and 5 show the results of estimating the equation separately for lower- and upper-division classes. The impact of beauty on instructional ratings, while statistically significant in both equations, is over twice as large in lower-division classes. Indeed, the same much bigger effects are found for two of the other variables that affected instructional ratings in the sample as a whole, whether the instructor is on the tenure track or is female. We might be tempted to conclude that class ratings by more mature students, and students who are learning beyond the introductory level in a subject, are less affected by factors such as beauty that are probably unrelated to the instructor's knowledge of the subject. Yet the impacts of being a minority faculty member or a non-native English speaker are just as large in the estimates for upper-division courses as in those for lower-division courses. It is unclear why the impacts of these variables among those in X are not attenuated in the more advanced courses. These estimates may imply the existence of discrimination by students in their evaluations, or they may result from shortfalls in the ability of those instructors to transmit knowledge or inspire students.

3.2. Robustness tests

One might be concerned that a host of statistical problems plagues the estimates shown in Table 3 and implies that our results are spurious. One difficulty is a potential measurement error: raters may be unable to distinguish physical attractiveness from good grooming and dress. Were this merely classical measurement error, we would have no difficulties. A subtle problem arises, however, if those who dress better, and whose photographs may thus be rated higher, are the same people who take care to be organized in class, to come to class on time, to hold their announced office hours, etc. What if our measure of beauty is merely a proxy for the general quality of the faculty member independent of his/her looks?

To account for this possibility we created an indicator equaling one for male faculty members who are wearing neckties in their pictures and for female faculty who are wearing a jacket and blouse. Formal pictures are on the websites of one-sixth of the faculty (weighted by numbers of students), and this indicator is added to a respecified version of the basic equation for which the results were shown in Column 1 of Table 3. The estimated impacts of this indicator and of composite standardized beauty are presented in the first row of Table 4. While instructors who present a formal picture do receive higher ratings, the inclusion of this additional measure reduces the estimated impact of beauty only slightly. The effect of composite standardized beauty remains quite large and highly significant statistically. We may conclude that the potential positive correlation of measurement error in the beauty ratings with unobservable determinants of teaching success does not generate serious biases in our estimates.

A related problem, also involving possibly non-classical measurement error, might arise if the more

Table 4
Alternative estimates of the relation between beauty and class ratings (lower- and upper-division classes)

		Variable				
		Composite standardized beauty	Formal dress	Black and white	Composite standardized beauty	
					Above mean	Below mean
1.	Photo bias (dress) ($N = 463$)	0.229 (0.047)	0.243 (0.088)			
2.	Photo bias (picture quality) ($N = 463$)	0.267 (0.063)		0.088 (0.106)		
3.	Photo bias (department) ($N = 414$)	0.236 (0.049)				
4.	Asymmetric beauty effect ($N = 463$)				0.237 (0.096)	−0.318 (0.133)
5.	Course fixed effects ($N = 157$)	0.177 (0.107)				

Note: The equations reported in rows 1–4 also include all the variables included in the basic equation in column 1 of Table 3. The equation reported in row 5 excludes variables in the vector Z.

concerned instructors were concerned enough about their pictures to include color rather than black-and-white photos on the websites. We classified the photographs along this criterion and again reestimated the basic equation. As the second row of Table 4 shows, there was almost no change in the parameter describing the relationship between composite standardized beauty and the evaluation. While the coefficient on the indicator variable "black-and-white" was small and statistically quite insignificant, it was somewhat surprisingly positive.[7]

Perhaps the most serious potential problem may result from a type of sample selectivity. Consider the following possibility. Among a group of people (a department), those who place their photographs on their websites will, until equilibrium in the game is reached, be better looking than those who do not present their photographs. They may also be people who are "go-getters" in other aspects of their lives, including their classroom teaching. If that is true, those instructors who are among the few in a department whose pictures are available will be better looking and be better instructors, while those from departments with all pictures available will on average be average looking and average instructors.

To examine this potential problem we reestimate the basic equation on the subsample of 84 faculty members, teaching 414 classes, in which an entire department's faculty's pictures are available. The results of estimating the basic equation over this slightly reduced sample are shown in the second row of Table 4. Compared to the basic estimate (0.275), accounting for this potential problem reduces the estimated impact of composite standardized beauty slightly and implies that a two-standard deviation change in beauty raises the course rating by 0.39 (three-fourths of a standard deviation in course ratings). Apparently this kind of selectivity matters a bit, but it does not vitiate the basic result.

The next possibility does not represent a potential bias in the basic results, but rather asks whether they are masking some additional sample information. There is some indication (Hamermesh & Biddle, 1994; Hamermesh, Meng, & Zhang, 2002) that the effect of beauty on earnings is asymmetric, with greater effects of bad than of good looks. Does this asymmetry carry over into its effects on productivity in college teaching? To examine this possibility we decompose the composite standardized beauty measure into positive and negative values and reestimate the basic equation allowing for asymmetry. The results are shown in the third row of Table 4. The effect on course ratings of looking better than average is slightly below and opposite in sign of the effect of looking worse than average.[8] There is only slight evidence of asymmetry in the impact of instructors' beauty on their course ratings.

Another potential issue is that courses may attract students with different attitudes toward beauty. These may be correlated with the instructional ratings that the students give and may also induce departmental administrators to assign courses to instructors based on their looks. Some courses may also generate different ratings

[7]Yet another potential difficulty is that the photographs may not all be equally current. Given that all had to be in electronic files, and given the strong evidence (Hatfield & Sprecher, 1986, pp. 282–3) that an individual's perceived beauty changes very slowly with age, even a correlation between the age of the photograph and an instructor's evaluation would cause at most a minimal bias in any estimates.

[8]The t-statistic on the hypothesis that they are equal and opposite in sign is 0.41. This may not contradict results indicating asymmetric effects of beauty on earnings. Many more individuals are rated above average in looks than are considered below average, so that the asymmetry might not exist if the beauty measure itself were symmetric, as it is by construction here.

depending on their difficulty, their level and other differences, and these may be correlated with the instructor's looks. The gender mix of students may differ among courses, and this too may affect the estimated impacts of beauty. To examine these possibilities we take advantage of the fact that 157 of the 463 classes in our sample are instructed by more than one faculty member over the 2 years of observation. These courses involve 54 different instructors (of the 94 in the sample). We reestimate the basic equation on this subsample adding course fixed effects. Thus any estimated effect of beauty will reflect within-course differences in the impact of looks on instructional ratings.

The results are presented in the final row of Table 4. The estimated impact of composite standardized beauty on class evaluations is somewhat smaller than in the other estimates, but still substantial. This is mostly due to sampling variability. Reestimating the basic equation of Table 3 over this reduced sample of 157 classes yields an impact of composite standardized beauty on instructional ratings of 0.190 (s.e. = 0.079).[9]

4. Conclusion and interpretations

The estimates leave little doubt that measures of perceived beauty have a substantial independent positive impact on instructional ratings by undergraduate students. We have accounted for a variety of possibly related correlates, and we have shown that the estimated impacts are robust to potential problems of selectivity, correlated measurement error and other difficulties. The question is whether these findings really mean that beauty itself makes instructors more productive in the classroom, or whether students are merely reacting to an irrelevant characteristic that differs among instructors.

The first issue is that our measure of beauty may merely be a proxy for a variety of related unmeasured characteristics that might positively affect instructional ratings. To the extent that these are positively correlated with beauty but not caused by it, our results overstate the impact of beauty. That we have held constant for as many course and instructor characteristics as we have should mitigate some concerns about this potential problem. If there is a characteristic that is caused by a person's physical appearance and that also generates higher instructional ratings, then failing to measure it (and excluding it from the regressions) is correct. For example, if good-looking instructors are more self-confident because their beauty previously generated better treatment by other people, and if their self-confidence makes them more appealing instructors, it is their beauty that is the ultimate determinant of (part of) their teaching success.

A second and more important issue is whether higher instructional ratings mean that the faculty member is a better teacher—is more productive in stimulating students' learning. The instructional ratings may putatively reflect productivity, but do they really do so? Discussions of this question among administrators and faculty members have proceeded since instructional evaluation was introduced, and we do not wish to add to the noise. Regardless of the evidence and of beliefs about this issue, however, instructional ratings are part of what universities use in their evaluations of faculty performance—in setting salaries, in determining promotion, and in awarding special recognition, such as teaching awards. Thus even if instructional ratings have little or nothing to do with actual teaching productivity, university administrators behave as if they believe that they do, and they link economic rewards to them. Thus the ratings are at least one of the proximately affected outcomes of beauty that in turn feed into labor-market outcomes.

The most important issue is what our results tell us about whether students are discriminating against ugly instructors or whether they really do learn less (assuming that instructional ratings reflect learning). For example, what if students simply pay more attention to good-looking instructors and learn more from them? We would argue that this is a productivity effect—we would claim that the instructors are better teachers. Others might (we think incorrectly) claim that the higher productivity arises from students' (society's) treating them differently from their worse-looking colleagues and is evidence of discrimination. Disentangling the effects of differential outcomes resulting from productivity differences and those resulting from discrimination is extremely difficult in all cases, as we believe this unusual illustration of the impact of beauty on a physical measure that is related to earnings illustrates.

The epigraph to this study may be correct—someone who does not qualify to be a supermodel might well go into teaching. Even in college teaching, however, our evidence demonstrates that a measure that is viewed as reflecting teaching productivity, whether it really does so or not, is also one that is enhanced by the instructor's pulchritude.

Acknowledgements

We thank William Becker, Jeff Biddle, Lawrence Kahn, Preston McAfee, Alex Minicozzi, Gerald Oettinger, a referee and participants at seminars at several universities for helpful suggestions.

[9] If we include a vector of indicators for departments in the basic equation in Table 3, we find a somewhat larger effect than here, although one that is still smaller than that in the basic equation.

References

Ambady, N., & Rosenthal, R. (1993). Half a minute: predicting teacher evaluations from thin slices of nonverbal behavior and physical attractiveness. *Journal of Personality and Social Psychology*, *64*, 431–441.

Becker, W., Watts, M. (1999). How departments of economics evaluate teaching. *Papers and proceedings 89*, American Economic Association (pp. 355–359).

Biddle, J., & Hamermesh, D. (1998). Beauty, productivity and discrimination: lawyers' looks and lucre. *Journal of Labor Economics*, *16*, 172–201.

Borjas, G. (2000). Foreign-born teaching assistants and the academic performance of undergraduates. *Papers and proceedings 90*, American Economic Association (pp. 344–349).

Buck, S., & Tiene, D. (1989). The impact of physical attractiveness, gender, and teaching philosophy on teacher evaluations. *Journal of Educational Research*, *82*, 172–177.

DeLorme, C., Hill, R. C., & Wood, N. (1979). Analysis of a quantitative method of determining faculty salaries. *Journal of Economic Education*, *11*, 20–25.

Feldman, K. (1993). College students' views of male and female college teachers: part II. Evidence from students' evaluations of their classroom teachers. *Research in Higher Education*, *34*, 151–211.

Fleisher, B., Hashimoto, M., & Weinberg, B. (2002). Foreign GTAs can be effective teachers of economics. *Journal of Economic Education*, *33*, 299–326.

Goebel, B., & Cashen, V. (1979). Age, sex and attractiveness as factors in student ratings of teachers: a developmental study. *Journal of Educational Psychology*, *71*, 646–653.

Hamermesh, D., & Biddle, J. (1994). Beauty and the labor market. *American Economic Review*, *84*, 1174–1194.

Hamermesh, D., Meng, X., & Zhang, J. (2002). Dress for success: does primping pay? *Labour Economics*, *9*, 361–373.

Hatfield, E., & Sprecher, S. (1986). *Mirror, Mirror* …. Albany, NY: State University of New York Press.

Katz, D. (1973). Faculty salaries, promotions, and productivity at a large university. *Papers and proceedings 63*, American Economic Association (pp. 469–477).

Kaun, D. (1984). Faculty advancement in a nontraditional university environment. *Industrial and Labor Relations Review*, *37*, 592–606.

Moore, W. J., Newman, R., & Turnbull, G. (1998). Do academic salaries decline with seniority? *Journal of Labor Economics*, *16*, 352–366.

Siegfried, J., White, K. (1973). Financial rewards to research and teaching: a case study of academic economists. *Papers and proceedings 63*, American Economic Association (pp. 309–315).