Economics Letters 181 (2019) 116–119

Contents lists available at ScienceDirect

Economics Letters

journal homepage: www.elsevier.com/locate/ecolet

Appearance Discrimination in Grading? — Evidence from Migrant Schools in China ☆

Qihui Chen [a], Xiaobing Wang [b], Qiran Zhao [a,*]

[a] College of Economics and Management, China Agricultural University, Beijing, 100083, China
[b] China Center for Agricultural Policy, Peking University, Beijing, 100871, China

HIGHLIGHTS

- Appearance of migrant-school students was rated by machine-learning programs.
- Students' appearance positively affects their teacher-graded test scores.
- Estimations control for key confounders such as cognitive and mental development.

ARTICLE INFO

Article history:
Received 5 February 2019
Received in revised form 28 March 2019
Accepted 20 April 2019
Available online 25 April 2019

JEL classification:
I2 J7

Keywords:
Discrimination
Appearance
Education
Migrant children
China

ABSTRACT

Using appearance scores created by facial-recognition and machine-learning programs that incorporate tens of thousands of individuals' appearance preferences, we find in China's migrant schools that students' appearance has a statistically significant and positive effect on their teachers' evaluation of their exam performance, even after netting out the influences of important confounders such as physical growth, cognitive ability, mental health status, family background, and school quality.

1. Introduction

At least since Becker (1957), economists have been studying labor-market discrimination, with respect to race, gender, religion, and other ascriptive characteristics. A particular form of discrimination, appearance discrimination, has attracted growing attention since the mid-1990s. Empirical findings from various countries suggest that physically attractive workers are often paid better than less attractive ones (Hamermesh and Biddle, 1994; Biddle and Hamermesh, 1998; Harper, 2000; Robins et al., 2011; Scholz and Sicinski, 2015).

However, these findings do not necessarily imply the existence of appearance discrimination. Firstly, popular measures of physical (un)attractiveness (e.g. overweight) may be correlated with unobserved personal traits (e.g. lack of perseverance and endurance) that affect job performance. Thus, the association between physical attractiveness and labor-market outcomes may simply reflect the influence of these traits. More subtly, physical attractiveness may have productive value itself, which renders it challenging to detect appearance discrimination. Physically attractive (e.g. tall and fit) people might be more competitive and confident (Fessler et al., 2010; Hensley, 1993; Martel and Biller, 1987) — thus more productive at work. Since it is difficult to measure to what extent these productive aspects are pecuniarily rewarded, isolating appearance discrimination from monetary returns to appearance is difficult. Worse yet, if supervisors exploit such difficulty in detection to practice appearance discrimination, deliberately inflating their evaluation of attractive employees' job performance, detecting appearance discrimination is even more difficult.

This study isolates appearance discrimination from monetary returns to appearance by targeting a particular population whose physical appearance has little economic value — students attending *migrant schools* in China — and examines how their appearance affects their supervisors' (teachers') evaluation of their

☆ This work was supported by the National Natural Science Foundation of China [grant numbers 71603261, 71673008, 71742002]; Humanities and Social Science Fund of Ministry of Education of China [grant numbers 16YJC880107, 18YJC790010].

* Corresponding author.
E-mail address: zhaoqiran@cau.edu.cn (Q. Zhao).

https://doi.org/10.1016/j.econlet.2019.04.024
0165-1765/

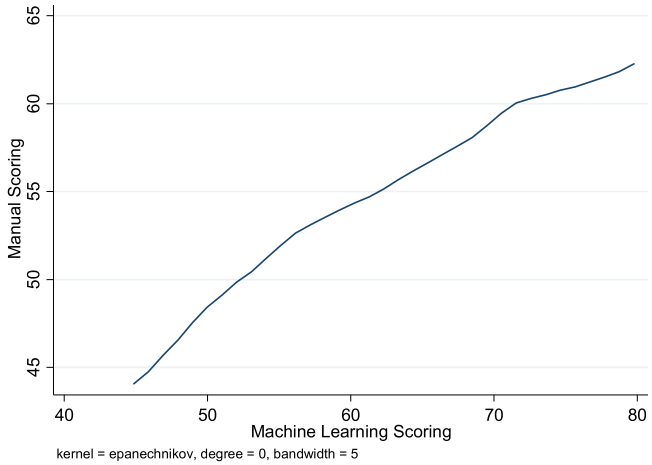

Fig. 1. Correlation between Machine- and Human-Rated Appearance Scores. Notes: the *x*-axis and *y*-axis present, respectively, the machine-rated appearance scores and human-rated scores (given by 40 researchers) for 100 randomly chosen photos.

Table 1
Associations between standardized appearance scores and teacher-graded test scores.

	Math	Chinese	English
Appearance	0.031*	0.035*	0.050***
	(0.015)	(0.018)	(0.016)
Original *p*-value	[0.0408]	[0.0521]	[0.0024]
Romano–Wolf *p*-value	[0.0677]	[0.0677]	[0.0040]
Controls	Yes	Yes	Yes
Observations	3356	3356	3356
R^2	0.190	0.331	0.295

Notes: Regressions that include the full set of covariates are reported in Appendix Table A.1. Standard errors are reported in parentheses, adjusted for within-school clustering.
Three hypotheses are being tested when computing the stepdown p-values (Romano and Wolf, 2005, 2016).
*p<0.10.
***p<0.01.

Table 2
Impacts of appearance on teacher-graded test scores around potential Grading–Lifting "targets".

A. Range of raw math scores:	45–54	55–64	65–74	75–84	85-94
Appearance	−0.009	0.129*	0.010	0.027	−0.008
	(0.041)	(0.044)	(0.044)	(0.018)	(0.012)
	[0.436]	[0.079]	[0.812]	[0.347]	[0.673]
Observations	109	178	305	594	1208
R^2	0.879	0.777	0.745	0.710	0.537
B. Range of raw Chinese scores:	45–54	55–64	65–74	75–84	85-94
Appearance	−0.012	0.013	−0.010	−0.001	0.002
	(0.027)	(0.011)	(0.009)	(0.005)	(0.004)
	[0.951]	[0.515]	[0.525]	[0.960]	[0.505]
Observations	89	217	334	628	1393
R^2	0.589	0.301	0.153	0.142	0.124
C. Range of raw English scores:	45–54	55–64	65–74	75–84	85-94
Appearance	−0.009	−0.012	0.010	0.006	−0.002
	(0.021)	(0.011)	(0.009)	(0.006)	(0.004)
	[0.782]	[0.158]	[0.426]	[0.624]	[0.446]
Observations	142	274	367	644	985
R^2	0.476	0.312	0.263	0.181	0.176

Notes: Regressions that include the full set of covariates are reported in Appendix Table A.1. Standard errors are reported in parentheses, adjusted for within-school clustering. Stepdown p-values (Romano and Wolf, 2005, 2016) with 250 bootstrapping replicates are reported in brackets. Fifteen hypotheses are being tested when computing the stepdown p-values.
*p<0.1.

job (academic) performance. The institutional setting pertaining to migrant schools in China renders their students an interesting population to study. While migrant schools adopt the same curriculum as regular schools to fulfill their role as education facilities, they are constructed to *temporarily* accommodate children of rural migrants working in the cities (Chen and Feng, 2013) — according to local regulations, migrant students (without legal residential permits) must return to their hometowns to pursue high school education (Wang et al., 2017). Such a setting predicts that teachers in migrant schools have little economic incentive to inflate students' test scores — doing so will not win them teaching awards; nor can it increase students' admission to high schools.

To obtain a direct measure of (rather than proxies for) appearance, we apply face-recognition and machine-learning techniques to create appearance scores for nearly 3500 migrant students based on their photos taken during our survey. Our regressions reveal a statistically significant and positive association between one's appearance and teacher-graded test scores, even after controlling for potential confounders reflecting students' physical, cognitive and mental development, family background, and school quality.

2. Data

Our data were collected through a school-based survey conducted in two major migration destination cities in China, Beijing in the north and Suzhou in the south, in June 2017. Among all migrant schools operating in these two cities, 30 representative schools in Beijing and 29 in Suzhou were chosen. In each chosen school, each class in third and class fourth grades was randomly chosen, yielding a sample of 3356 students. Besides a standard survey on students' personal and family characteristics, we also obtained their test scores from the previous semester from administrative records. Tests were also administered to assess their physical development (e.g. height and weight), cognitive ability, and mental health status (e.g. self-esteem and depression),[1] the distributions of which are summarized in column (1) of Appendix Table A.1.

Most importantly, we took photos of all sampled students (with their consent) and hired a face++ company to design a

[1] Students' cognitive ability was assessed by Raven's Standard Progressive Matrices test (Raven et al., 2004). Their self-esteem was assessed by Rosenberg's (1965) widely-adopted RSES scale, modified to fit the Chinese context. Depression was assessed by a 6-item scale adapted from the widely-used CES-D scale (Radloff, 1991).

Table A.1
Summary statistics and associations between standardized appearance scores and teacher-graded test scores.

	Mean [Std. Dev.]	Math		Chinese		English	
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Appearance	0.00	0.037**	0.031**	0.042**	0.035*	0.063***	0.050***
	[1.00]	(0.015)	(0.015)	(0.017)	(0.018)	(0.018)	(0.016)
Boy	0.55	−0.103***	−0.065*	−0.321***	−0.290***	−0.367***	−0.330***
	[0.50]	(0.038)	(0.035)	(0.027)	(0.028)	(0.029)	(0.027)
Age (months)	127.4	−0.002	0.004**	−0.001	0.003	−0.004	−0.000
	[10.44]	(0.002)	(0.002)	(0.002)	(0.002)	(0.003)	(0.002)
Preschool attendance	0.89	0.293***	0.170***	0.185***	0.106**	0.239***	0.164***
	[0.31]	(0.064)	(0.051)	(0.056)	(0.050)	(0.054)	(0.051)
Sibship size	1.41	−0.101**	−0.042	−0.100***	−0.060**	−0.085**	−0.048
	[0.59]	(0.040)	(0.035)	(0.033)	(0.030)	(0.032)	(0.029)
Father's education (years)	9.33	0.018***	0.012*	0.016***	0.011**	0.010	0.006
	[2.72]	(0.006)	(0.007)	(0.004)	(0.005)	(0.006)	(0.007)
Mother's education (years)	8.74	0.017***	0.015***	0.006	0.004	0.004	0.002
	[3.18]	(0.005)	(0.005)	(0.005)	(0.005)	(0.006)	(0.005)
Household asset	0.00	0.086***	0.057***	0.061***	0.040***	0.076***	0.058***
	[1.00]	(0.018)	(0.017)	(0.013)	(0.012)	(0.013)	(0.012)
Height (cm)	138.2		0.000		−0.001		−0.001
	[7.83]		(0.002)		(0.002)		(0.002)
Weight (BMI)	17.41		0.005		0.002		−0.003
	[3.12]		(0.005)		(0.005)		(0.005)
Raven's test scores	90.33		0.025***		0.015***		0.014***
	[14.16]		(0.002)		(0.001)		(0.001)
Self-esteem	46.10		0.016***		0.014***		0.011***
	[8.10]		(0.002)		(0.002)		(0.002)
Depression	0.30		−0.115***		−0.157***		−0.181***
	[0.46]		(0.041)		(0.034)		(0.034)
School fixed effects		Yes	Yes	Yes	Yes	Yes	Yes
Observations	3356	3356	3356	3356	3356	3356	3356
R^2		0.039	0.190	0.239	0.331	0.217	0.295

Notes: Standard errors in parentheses, adjusted for within-school clustering.
*p<0.10.
**p<0.05.
***p<0.01.

facial-recognition and appearance-rating program. Built upon Neural Networks principles, the program was trained using appearance scores of millions of eastern faces rated by tens of thousands of individuals.[2] Our machine-rated appearance measure differs from other existing measures (e.g. Talamas et al., 2016), in that our measure incorporates tens of thousands of individuals' appearance preferences, whereas other measures typically rely on only a few enumerators' preferences. Indeed, assessing the correlation between our machine-rated scores and the average ratings by 40 fellow researchers for 100 randomly chosen photos reveals that our machine-rated scores are very consistent with human evaluations (Fig. 1).

3. Method

Since it is virtually impossible to experimentally change one's appearance in our setting, our interest is centered on the partial association between one's appearance (*Appearance*) and teacher-graded test scores (*Score*), netting out the influences of physical, cognitive and mental development. Formally, we estimate the following model:

$$Score = \beta + \beta_A \times Appearance + \boldsymbol{X\beta} + \varepsilon \quad (1)$$

where the set of control variables **X** include child characteristics (i.e. gender, age, height, weight, Raven's test scores, preschool attendance, self-esteem and depression status), family characteristics (i.e. parental education, family asset holding and sibship size), and school fixed effects. For ease of interpretation, all scores are standardized to have zero mean and unity standard deviation.

4. Results

Table 1, columns 1–3, reports results of estimating Eq. (1), respectively, for teacher-graded math, Chinese and English scores. All regressions include the full set of control variables reported in Appendix Table A.1. For all subjects, *Appearance* positively (and statistically significantly, at least marginally so) predicts teachers' evaluation of students' exam performance, even after controlling for important confounding factors such as cognitive ability, stature, mental health, family background and school quality. To further account for the fact that multiple hypotheses are being tested in these regressions, we implement Romano and Wolf's (2005, 2016) stepdown procedure to derive adjusted p-values for the appearance effects (Table 1). The stepdown p-values (with 250 bootstrapping replicates) are somewhat larger than the original ones but still suggest statistically significant appearance effects, at least at the 0.0677 level.[3] Also, the predictive power of appearance is larger for language scores (especially English scores) than for math scores, presumably because solutions to math problems are in general more objective, leaving not much room for grade-lifting.

Where, then, is room for grade-lifting in math grading? One possibility is that teachers tend to lift better-looking students' scores around the passing mark (at 60). Panel A of Table 2 reports coefficients of *Appearance* estimated in a consecutive set of 10-point intervals around the 50, 60, 70, 80 and 90 marks, i.e. potential "targets" for grade-lifting. While the coefficients in

[2] Trained enumerators checked the quality of photos carefully during the survey to ensure that students' photos can be identified by the program: In photo-taking, a student sitting against the white background should face the camera and cannot look up or down or squint in the daylight. His/her face needs to occupy more than 50% of the area of the entire photo.

[3] We thank the Editor of the Journal, Professor Costas Meghir for suggesting this test.

the middle range (55–84) are all positive, only the one around 60 is statistically significant (original p-value < 0.01; Romano–Wolf p-value $= 0.079$); it is also much larger than estimates around other "targets". Also consistent with the "grade-lifting around 60" hypothesis for math tests, as the intervals shrink from 10 to 8 and to 6 points, the coefficient increases from 0.129 to 0.137 and then to 0.159, all being statistically significant (at least marginally so).[4] In contrast, no such pattern was detected for either Chinese (Panel B) or English scores (Panel C), which suggests that rather than performing grade-lifting around specific "targets", language teachers might perform grade-lifting along the entire test-score spectrum.

5. Concluding remarks

Controlling for a set of important confounding factors, our analysis based on machine-rated appearance scores provides evidence that appearance discrimination exists even before one enters the labor market. While its impacts on grading are quantitatively small, it may generate a series of undesirable consequences. It may reduce better-looking students' effort, thereby undermining their cognitive development. It may also create an atmosphere of unfairness among students, likely undermining other students' cognitive development as well. More seriously, once children accept appearance discrimination as the norm in the society, their willingness to fight against it in the future labor market may be greatly reduced.

Appendix

See Table A.1.

References

Becker, G.S., 1957. The Economics of Discrimination. University of Chicago Press, Chicago.
Biddle, J.E., Hamermesh, D.S., 1998. Beauty, productivity, and discrimination: Lawyers' looks and lucre. J. Labor Econ. 16, 172–201.
Chen, Y., Feng, S., 2013. Access to public schools and the education of migrant children in China. China Econ. Rev. 26, 75–88.
Fessler, D., Gneezy, U., List, J., Hoffman, M., 2010. Height and Competitiveness. University of California—San Diego Working Paper.
Hamermesh, D.S., Biddle, J.E., 1994. Beauty and the labor market. Amer. Econ. Rev. 84, 1174.
Harper, B., 2000. Beauty, stature and the labour market: A British cohort study. Oxf. Bull. Econ. Stat. 62, 771–800.
Hensley, W.E., 1993. Height as a measure of success in academe. Psychol.: J. Hum. Behav. 30, 40–46.
Martel, L.F., Biller, H.B., 1987. Stature and Stigma: The Biopsychosocial Development of Short Males. Lexington Books.
Radloff, L.S., 1991. The use of the center for epidemiologic studies depression scale in adolescents and young adults. J. Youth Adolesc. 20, 149–166.
Raven, J., Raven, J.C., Court, J.H., 2004. Manual for Raven'S Progressive Matrices and Vocabulary Scales. Harcourt Assessment, San Antonio.
Robins, P.K., Homer, J.F., French, M.T., 2011. Beauty and the labor market: accounting for the additional effects of personality and grooming. Labour 25, 228–251.
Romano, J.P., Wolf, M., 2005. Exact and approximate stepdown methods for multiple hypothesis testing. J. Amer. Statist. Assoc. 100, 94–108.
Romano, J.P., Wolf, M., 2016. Efficient computation of adjusted p-values for resampling-based stepdown multiple testing. Statist. Probab. Lett. 113, 38–40.
Rosenberg, M., 1965. Society and the Adolescent Self-Image. Princeton University Press, Princeton.
Scholz, J.K., Sicinski, K., 2015. Facial attractiveness and lifetime earnings: Evidence from a cohort study. Rev. Econ. Stat. 97, 14–28.
Talamas, S., Mavor, K., Perrett, D., 2016. Blinded by beauty: attractiveness bias and accurate perceptions of academic performance. PLoS One 11 (2), e0148284.
Wang, X., Luo, R., Zhang, L., Rozelle, S., 2017. The education gap of China's migrant children and rural counterparts. J. Dev. Stud. 53, 1865–1881.

[4] Results are not shown in the table but available upon request.