Gender Bias in Assessments of Teacher Performance

BY SABRIN BEG, ANNE FITZPATRICK, ADRIENNE M. LUCAS[1]

Professional advancement often depends on both a manager's assessment and an employee's own self-assessment of their productivity. Both assessments are often complicated due to a lack of objective outcomes and measures. Instead, managers may instead rely upon subjective assessments that reflect pre-existing biases based on characteristics like gender. Lacking objective data, and potentially receiving biased assessments, female employees may not develop accurate assessments of their own self-effectiveness. In this paper we compared managers' and workers' subjective assessments, testing whether that relationship varies by workers' genders. We then compared managers' assessments to an objective, output-based measure of productivity and asked managers to assess hypothetical people, randomly varying the gendered name of the person.

While performance reviews occur across almost all sectors of the economy, the education sector is a particularly interesting one to study. Identifying effective teachers, including promoting those who might be effective principals, is crucial to improving school quality and student learning. Education also contains observable metrics: a teacher's effectiveness can be measured objectively with students' test score gains (Bau and Das 2020). However, in developing countries, principals may lack such data, leading to assessments that could include both classical measurement error and systematic bias based on teachers' characteristics that are unrelated to their actual effectiveness (Harris and Sass 2014). Lacking data, teachers themselves may also be unaware of their true

[1] Beg: University of Delaware, 418 Purnell Hall, Newark, DE 19716 (email: sbeg@udel.edu). Fitzpatrick: University of Massachusetts Boston, 100 Morrissey Blvd., Wheatley Hall 5-025, Boston, MA 02125 (email: anne.fitzpatrick@umb.edu). Lucas: University of Delaware, 419 Purnell Hall, Newark, DE 19716, NBER, J-PAL, and CGD (email: alucas@udel.edu). We thank the World Bank Strategic Impact Evaluation Fund (SIEF), Abdul Latif Jameel Poverty Action Lab (J-PAL), and the United Nations Children's Fund for generous funding of this project. For exceptional project management and research assistance in Ghana we thank Henry Atimone, Renaud Comba, and Edward Tsinigo. This project would not have been possible without the dedication of the entire IPA Ghana STARS policy and field teams, especially Joyce Jumpah and Bridget Konadu Gyamfi. We thank UNICEF and Ghana Education Services for their partnership and our respondents for their cooperation. For useful comments and suggestions we thank Heather Sarsons and Petra Todd.

productivity. As a result, neither teacher effectiveness nor student learning might reach their full potential.

Formally, we compared both a Ghanaian school principal's assessment of teacher effectiveness and a teacher's own self-assessment of effectiveness relative to an objective measure based on student test score increases, i.e. value-added, testing for systematic differences on the basis of teacher gender. Ghanaian education is an especially salient location to study issues of gender bias as in our sample of primary schools a majority of both teachers and principals were male and gender bias is more prevalent in male-dominated fields (Blau and Kahn 2017).

When asked to compare themselves to other teachers at similar schools, female and male teachers were equally likely to rate themselves as a more effective than other teachers, while principals were about 11 percentage points less likely to rate a female teacher as more effective. By contrast, female teachers were objectively more effective—student learning was higher for female teachers. When ranked by objective effectiveness, principals' subjective assessment of male teachers exceeded that of female teachers at all points in the objective distribution. The least effective male teacher was still subjectively assessed as more effective than the most effective female teacher. We augment our results with a survey experiment. We described scenarios to each principal involving a hypothetical teacher, principal, or principal supervisor, each randomly assigned either a female or male name. Principals assessed women as 0.10 standard deviations (SDs) less effective than if the same action was taken by a man.

While we cannot rule out all other explanations, our results indicate that some of the difference in principal evaluation of teachers was likely due to gender bias, providing one explanation for why the gender ratio decreases within the education sector hierarchy—in our sample, females were 47% of primary school students but only 8% of principal supervisors.

Overall, these results contribute additional evidence on the presence of gender bias in subjective assessments (Goldin and Rouse 2000; Blau and Devaro 2007; Sarsons 2017; Sarsons et al. forthcoming) and implicit bias in evaluation of equally qualified candidates (Moss-Racusin et al. 2012). We add evidence on additional barriers that women face within labor markets in lower income countries due to the paucity of objective data on performance beyond the barriers they face to enter labor markets (Jayachandran 2020).

I. Background and Setting

Four characteristics make the Ghanaian primary education context particularly well suited to evaluating gender bias in subjective assessments and eliminate some common concerns about value added measures (Rothstein 2009). First, principals have very little objective information about teacher performance as no standardized testing occurs during primary school, grades 1 to 6. Second, teachers are classroom teachers, teaching all subjects, and most schools in our sample only have one section of each grade, limiting principal discretion in assigning students to teachers. Third, principals cannot hire or fire teachers and teachers do not directly apply to a specific school, limiting teacher sorting across schools.[2] Fourth, school principals play a key role in creating a productive work environment, contributing to teacher retention in the sector, and influence which teachers become principals or advance to other leadership positions in the schooling sector.

II. Conceptual Framework and Empirical Strategy

Understanding whether bias is present in performance assessment is often confounded by the lack of an objective assessment of effectiveness. Our subjective measures relied on survey responses from principals and teachers. A disagreement in assessment between an employee and a supervisor is not necessarily a sign of bias. In the education sector, student test scores are one

[2] Ghana Education Services (GES) hires teachers and assigns each one to a particular school. To change schools, teachers apply directly to GES to leave their existing school. Transfers are at the sole discretion of GES.

objective measure of effectiveness—we calculated a value-added measure for each teacher based on student test score gains that isolated the effect of the teacher plus the overall school environment net of any other pre-existing contributions to the students' test scores (see Section III).

To test the correlation between teacher gender and our three measures of assessment (self, principal's, and objective) we estimated the following regression

$$Y_i = \alpha + \beta F_i + \boldsymbol{X}_i{}'\boldsymbol{\gamma} + \varepsilon_i \quad (1)$$

where Y_i is the measure of assessment for teacher *i,* F_i is an indicator equal to 1 if teacher *i* is a woman, $\boldsymbol{X_i'}$ is a vector of teacher-specific control variables, plus principal demographics when the outcome of interest is the principal's assessment, and ε_i are standard errors clustered at the school level.[3] The primary coefficient of interest, β, tests whether after controlling for other observable covariates, female and male teachers differ in perceived (from the subjective measures) or actual (from the objective measure) effectiveness.

We further tested whether the relationship between objective and subjective effectiveness varied by gender by ranking teachers by their objective effectiveness and plotting that relative to their principals' assessment of their effectiveness in a non-parametric way, separately by gender.

Finally, we provided respondents with hypothetical vignettes with a randomly assigned gender name to the person in the vignettes and use a modified Equation 1 to test for gender bias in the assessment of the effectiveness of hypothetical people,

$$Y_{iv} = \alpha + \beta F_{iv} + \gamma_i + \delta_v + \varepsilon_{iv} \quad (2)$$

where Y_{iv} is the assessment of person *i* in vignette *v*, F_{iv} is an indicator equal to 1 if the person in the vignette had a female name, γ_i are respondent fixed effects, δ_v are vignette fixed effects, and

[3]For the subjective assessments we include class size, grade, STARS treatment status (assigned at the school level) and teacher age, age squared, years of experience, years of experience-squared, and indicator variables for education degree level as additional controls. The first three controls are already removed in the calculation of the objective measure. The principal demographics are age, age-squared, experience, experience squared, and gender. All data were collected as part of the STARS randomized controlled trial. For more information see Beg, Fitzpatrick, and Lucas (2020).

ε_{iv} is a standard error clustered at the respondent level. As with Equation 1, the primary coefficient of interest is β, the extent to which hypothetical women and men are assessed differently when performing their assigned job duties equally.

III. Data

The data are from the Strengthening Teacher Accountability to Reach All Students (STARS) project, collected from 210 schools in 20 districts across all regions of Ghana.[4] We focused on teachers of grades 5 and 6, the final two years of primary school. We surveyed teachers, principals, and grade 4 and 5 students and invigilated assessments in math and English (the language of instruction) at the end of the 2017-2018 academic year. One year later, we returned to the same schools, again surveying the teachers and principals and assessing the students in math and English. As this was not a high stakes exam, teachers had no incentive to discourage or encourage specific students from taking the tests.

Demographics

Females were a minority of students, teachers, and principals, with the gender ratio becoming more skewed through each layer. About 47 percent of students, 22 percent of teachers, and 15 percent of principals were female. On average female teachers were 29 years old, with 4.6 years of experience, 2 years at the current school. Male teachers were 2 years older with an additional year of experience overall and at the current school. We control for these differences in our analysis.

Subjective Assessments

In both rounds of data collection we asked each principal to assess each of their teachers who would likely be or were currently teaching grades 5 and 6. Principals were asked whether a particular teacher was much less effective, less effective, as effective, more effective, or much more

[4] Full details on the sample selection appear in Beg, Fitzpatrick, and Lucas (2020).

effective relative to other teachers at similar schools. For ease of interpretation, we collapse this scale into a binary measure of whether the principal viewed the teacher as at least more effective.[5]

Objective Assessments

To calculate our measure of objective effectiveness we compared the test score gains of students across the two achievement rounds, controlling for student characteristics, following value-added method of Chetty et al. (2014) as much as possible given our data constraints. Because we only have one year of data, the measure includes both teacher and school value added.[6] We converted raw student math and English test scores into a latent ability measure using Item Response Theory (IRT), then estimated the following regression

$$ELscore_{ist} = \beta BLscore_{ist} + T_t + \boldsymbol{X}_{ist}'\boldsymbol{\gamma} + \varepsilon_{ist} \quad (3)$$

with one observation per subject by student where *ELscore* is the score at endline for student *i* in subject *s* taught by teacher *t*, *BLscore* is the same student's baseline test score, T_t are a series of teacher-specific fixed effects, $\boldsymbol{X}_{ist}$ are student-specific covariates (gender, grade fixed effect, subject fixed effect, age, and age-squared), grade-level covariates (the number of students per grade), and treatment status from the STARS project assigned at the school level. After controlling for these student and school observables, the coefficients on T_t are the sum of both the school's and teacher's objective effectiveness—the degree to which we can attribute test score gains to a teacher and school-specific input.

Hypothetical Assessments

[5] Results were similar using an ordered probit or a linear scale. To maximize our sample size and because the subjective assessments were largely consistent over time for teachers for whom we have two assessments, we use the subjective assessment from the endline data.

[6] A school fixed effect is an alternative control, but does not fit with how the effectiveness question was posed to the head teacher—asking him to compare teachers to other teachers in similar schools not to the other teachers in the same school. Further, using a school fixed effect strains our data, as it uses variation in schools that have both male and female teachers. As women are under-represented, that is only 27 percent (55) of the schools.

As a final measure of potential bias, we randomly allocated a male or a female name to each of 5 vignettes about a teacher, principal, or school supervisor performing their duties with varying degrees of effectiveness, asking the respondent to judge how effective they thought the person was on a scale from 1 (much less effective) to 5 (much more effective) than others in a similar position. To allow for comparability across vignettes, we standardize each response relative to the vignette mean and standard deviation.

IV. Results

Table 1 contains our estimates of Equation 1 with the three measures of effectiveness as the dependent variables. Female and male teachers were equally likely to assess themselves as at least more effective than other teachers at similar schools (column 1). In contrast, principals were about 11 percentage points less likely to assess female teachers this highly relative to male teachers (column 2). Therefore, female teachers showed relatively more confidence in their ability than principals did. In column 3, we test for gender differences in the objective measure of effectiveness based on student test scores and find that female teachers had on average 0.28SD higher effectiveness than their male peers. [7]

[Insert Table 1 Here]

Given our data limitations we cannot know whether female teachers are more effective or more likely to be in schools that are particularly effective.[8] Two factors indicate the former but we cannot dismiss the latter. First, principals generally rated teachers with higher objective assessment scores more highly, which would be unlikely if test score improvements were only the effect of schools. Second, teachers have limited ability to sort between schools as assignments are made

[7] This finding is robust to limiting the sample to schools with only one section of each grade and controlling for time-invariant school characteristics.

[8] As we are using student score gains to measure value added, the effect of the school characteristics would have to be affecting the slope of learning and not just the level.

centrally. Nevertheless, as teachers are assigned arbitrarily but not strictly randomly, female teachers could be requesting transfers away from lower quality schools or female teachers could be making school-level value added higher. Regardless, the data indicate that women were likely not less effective than their male peers.

In Figure 1 we examine the gender effect nonparametrically as a function of objective effectiveness. First, we rank teachers by objective assessment, creating percentiles of average test score gains. Then, we estimate the relationship between this percentile and residualized principal assessments using local linear regressions for each gender. The black dash-and-dot line is for male teachers and the blue dashed line is for female teachers. In both cases, the lines are mostly upwards sloping—within each gender, principals assessed teachers with higher objective effectiveness as more effective—providing support for an interpretation of the objective effectiveness capturing teacher effectiveness and not school effectiveness.[9] The differences between the genders are stark—across all percentiles, principals assessed male teachers as more effective than female teachers. The least effective male teacher was assessed as more effective than the most effective female teacher.

[Insert Figure 1 Here]

Test score gains are only one margin on which principals could be assessing teachers' overall effectiveness. In this context they are likely of primary importance to the principal's objective function—the only data that are collected on school "quality" are school exit exams at the end of junior high school and biennial regionally representative National Education Assessment exams in grades 4 and 6, and parents are highly focused on exam scores as they determine students' admissions and placement in senior high schools (Ajayi, Friedman, and Lucas 2017; Gilligan et al. forthcoming). Nevertheless, we offer an alternative value added calculation that punishes teachers for absent students, another margin on which they might be judged by principals, by giving absent

[9] The downward slope between 0 and 0.4 for women is because very few women are in that part of the distribution.

students a score of 0 on the follow-up exam. Our results are robust to this adjustment. Additionally, students assessed female teachers more highly, reporting that female teachers were 4.7 percentage points more likely to always give extra help (p-value of 0.058, male teacher average was 48%).

Principals further demonstrated evidence of bias against women in their hypothetical assessments. Principals rated individuals 0.12 standard deviations less effective when they had a female versus male name (column 4). Teachers assessed hypothetical people similarly regardless of the name's gender.

V. Discussion and Conclusions

Based on data collected on grade 5 and 6 classroom teachers in Ghana we find that female teachers assessed themselves as equally effective to their male peers while their principals assessed them as less effective. Using a measure of objective assessment based on student test scores, female teachers were more effective than their male peers or they taught at schools that were more effective.

From a non-parametric plot of objective versus subjective teacher effectiveness, principals were able to recognize effective teachers—principals assessed teachers with higher objective assessments as more effective. Regardless of their level of objective effectiveness, principals of female teachers assessed them as less effective than principals of male teachers. The principal assessment of the least objectively effective male teacher was higher than that of the most effective female teacher. Further, principals assess the effectiveness of hypothetical people as lower if the hypothetical person has a female instead of male name.

Beyond a general distaste for unfounded disparate treatment, bias against females may be especially detrimental to both teacher and student outcomes resulting in increased turnover (Cobbold 2015), lack of female role models for students (Dee 2004, Dee 2007), decreased test scores (Rockoff et al. 2012), and limited promotion opportunities (Cullen and Perez-Truglia 2020)

contributing to the underrepresentation of women in the management ranks of education—women are 22 percent of teachers, 15 percent of principals, but only 8 percent of school supervisors.

References

Ajayi, K. F., W. H. Friedman, and A. M. Lucas. 2017. "The Importance of Information Targeting for School Choice." *American Economic Review,* 107 (5): 638-43.

Beg, S., Fitzpatrick, A., and Lucas, A. M. 2020. "Successful Interventions at Scale: The Importance of Managers." mimeo.

Bau, N., and J. Das. 2020. "Teacher Value Added in a Low-Income Country." *American Economic Journal: Economic Policy*, 12 (1): 62-96.

Blau, F.D. and DeVaro, J., 2007. "New evidence on gender differences in promotion rates: An empirical analysis of a sample of new hires." *Industrial Relations: A Journal of Economy and Society*, *46*(3): 511-550.

Blau, F.D. and Kahn, L.M., 2017. "The gender wage gap: Extent, trends, and explanations." *Journal of Economic Literature*, *55*(3), pp.789-865.

Chetty, Raj, John N. Friedman, and Jonah E. Rockoff. 2014. "Measuring the Impacts of Teachers I: Evaluating Bias in Teacher Value-Added Estimates." *American Economic Review*, 104 (9): 2593-2632.

Cobbold, C., 2015. "Solving the Teacher Shortage Problem in Ghana: Critical Perspectives for Understanding the Issues." *Journal of Education and Practice*, *6*(9): 71-79.

Cullen, Z. B., & Perez-Truglia, R. (2020). "The Old Boys' Club: Schmoozing and the Gender Gap." National Bureau of Economic Research Working Paper No. w26530

Dee, T.S., 2004. "Teachers, race, and student achievement in a randomized experiment." *Review of economics and statistics*, *86*(1): 195-210.

Dee, T.S., 2007. Teachers and the gender gaps in student achievement. *Journal of Human resources*, *42*(3): 528-554.

Gilligan, D. O., N. Karachiwalla, I. Kasirye, A. M. Lucas, and D. Neal. Forthcoming. Educator Incentives and Educational Triage in Rural Primary Schools. *Journal of Human Resources.*

Goldin, C. and Rouse, C., 2000. Orchestrating impartiality: The impact of" blind" auditions on female musicians. *American economic review*, *90*(4): 715-741.

Harris, D.N. and Sass, T.R., 2014. Skills, productivity and the evaluation of teacher performance. *Economics of Education Review*, *40*: 183-204.

Jayachandran, S., 2020. Social Norms as a Barrier to Women's Employment in Developing Countries, mimeo.

Moss-Racusin, C.A., Dovidio, J.F., Brescoll, V.L., Graham, M.J. and Handelsman, J. 2012. "Faculty's subtle gender biases favor male students." *Proceedings of the National Academy of Sciences* 109(41): 16474-16479

Sarsons, H. 2017. "Recognition for Group Work: Gender Differences in Academia." *American Economic Review*, 107(5): 141-145.

Sarsons, H., Gërxhani, K., Reuben, E., and Schram, A. Forthcoming. "Gender Differences in Recognition for Group Work," *Journal of Political Economy*.

Rockoff, J. E., D. O. Staiger, T. J. Kane, and E. S. Taylor. 2012. Information and Employee Evaluation: Evidence from a Randomized Intervention in Public Schools. *American Economic Review*, 102 (7): 3184-3213.

Rothstein, J., 2009. Student sorting and bias in value-added estimation: Selection on observables and unobservables. *Education finance and policy*, *4*(4), pp.537-571.

TABLE 1— TEACHER EFFECTIVENESS

	More Effective or Much More Effective		Objective Assessment	Principal Assessment of Hypothetical Person
	Self- Assessment	Head Teacher Assessment		
	(1)	(2)	(3)	(4)
Female Teacher or Name	-0.0047	-0.109*	0.277**	-0.118*
	(0.0544)	(0.0633)	(0.135)	(0.070)
Observations	370	365	370	1044
R-squared	0.05	0.12	0.04	0.33
Male Teacher Average	0.70	0.60	-0.05	0.04

Notes: * significant at 10%; ** significant at 5%; *** significant at 1%. Standard errors clustered at the school level appear in parenthesis. Columns 1-3: Actual teachers. Columns 1 and 2: Linear probability models. Include controls for STARS treatment status, class size, grade, and teacher gender, age, age-squared, years of experience, years of experience-squared, and indicator variables for education degree level. Column 2: Additional controls include principal gender, age, age-squared, years of experience, and years of experience-squared. Column 3: Outcome in standard deviations, includes controls for teacher grade, age, age squared, years of experience, years of experience-squared, and indicator variables for education degree level. Column 4: Assessment of a hypothetical person whose gendered name was randomly assigned. Outcome in standard deviations of effectiveness. Controls are respondent and vignette fixed effects.

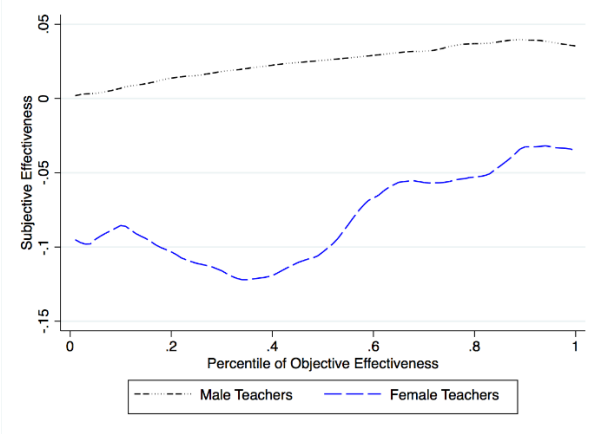

FIGURE 1. OBJECTIVE VS. SUBJECTIVE EFFECTIVENESS BY TEACHER GENDER

Note: The objective effectiveness is calculated based on student test score gains. The subjective effectiveness is whether the principal assessed the teacher as more effective or much more effective than other teachers in similar schools.