Online Appendices:

The impacts of performance pay on teacher effectiveness and retention:

Does teacher gender matter?

Andrew J. Hill [a]

Montana State University

Daniel B. Jones [b]

University of Pittsburgh

Abstract. Teacher performance pay is increasingly common in the United States. We assess the "incentive effects" of performance pay – the change in behavior of teachers present before and after a reform – with a focus on whether male and female teachers respond differently. Evaluating three performance pay programs in North Carolina, we find clear evidence of a gender difference: while male teachers' value-added remains flat before and after the introduction of performance pay, the value-added of female teachers declines. We also document suggestive evidence of a gender difference in retention, with men more likely to remain in schools with performance pay.

[a] Andrew Hill is an Assistant Professor in the Department of Agricultural Economics and Economics at Montana State University. Email: andrew.hill6@montana.edu.
[b] Daniel Jones is an Assistant Professor in the Graduate School of Public and International Affairs at the University of Pittsburgh. Email: dbj10@pitt.edu.

Appendix A: Student-teacher matching algorithm

We use counts of students in a set of grade and gender-by-race bins to measure classroom demographic composition. Grade is 9^{th}, 10^{th}, 11^{th} or 12^{th}, gender is male or female, and race is white, black or Hispanic, resulting in four grade bins and six gender-by-race bins. With the addition of total student count (reported separately to the other characteristics in the classroom-level file), there are a total of 11 dimensions used for describing classroom composition.

Matches on classroom demographic composition from the two sources are not expected to be perfect for a variety of reasons. First, the reported and constructed measures of classroom demographic composition are from different points in time; the classroom-level information is from the beginning of a course, while the student-level EOC file reflects composition at the end of a course (when students take the EOC test). Students may change classrooms or schools during this period. Second, some students may take the course but not write the EOC test (if they are absent on test day, for example). And, third, it is possible that classroom demographics from both sources are simply measured with error. As a result, we use the following algorithm to obtain course-specific student-teacher matches: (Matched classes are set aside after each step.)

1. In schools with only one teacher of the relevant course (in a given year), students and the teacher are matched. For example, in a school with only one teacher of Algebra I, all students writing the Algebra I EOC test are matched to this teacher. (13 percent of matched students are linked to their EOC teachers in this step.)
2. When reported classroom demographic composition (from the classroom-level file) perfectly matches constructed composition (from the student-level file) in all 11 categories, students from the student-level file are matched to teachers from the classroom-level file. (31 percent)

3. “Total student count” is excluded from the measure of classroom demographic composition (for this step and future steps). This is because it is the sum of students in either the grade or race-gender bins, so would exaggerate errors if the counts in any of these bins are incorrect. When reported composition perfectly matches constructed composition in the remaining 10 categories, students and teachers are matched. (<1 percent)
4. When reported composition perfectly matches constructed composition in 9 out of the 10 categories, students and teachers are matched if this match is unique and the deviation in the unmatched category is less than 2. In other words, students and teachers are matched if there is only a small mismatch on only one dimension of classroom demographic composition. If one classroom in the student-level file matches with multiple classrooms in the classroom-level file, students and teachers are matched to the classroom for which the deviation in the unmatched category is smallest (provided it is less than 2). If there are multiple matches with the same smallest deviation in the unmatched category, the classroom of students from the student-level file is dropped. (5 percent)
5. Repeat the above step, but link students and teachers with perfect matches on 8 out of 10 categories, and keep matches if the sum of deviations in the unmatched categories is less than 4. (26 percent)
6. The final steps in the algorithm use a fuzzy algorithm based on an overall distance measure: the sum of the absolute value of deviations in the 10 categories. Beginning with the constructed composition from the student-level file, find the best match in the classroom-level file, dropping classrooms from the student-level file with multiple best matches. Given that a classroom from the classroom-level file may be matched to

multiple classrooms in the student-level file, for every classroom in the classroom-level file, only keep the match with the smallest distance measure to ensure mutual best matching. Repeat the above step after setting aside the matches from the first iteration of the fuzzy algorithm. (25 percent)

Appendix B: VAM validity tests

We show in this section that the teacher VAMs we estimate in this paper (and use as a primary outcome variable) are not biased by the selection of students to teachers. We do so by performing two tests proposed by Chetty et al. (2014a; 2014b). First, we consider selection on observables. A given teacher's time-varying value-added measure should be (mechanically) correlated with the actual test score gains experienced by students in the teacher's class. Column 1 of Appendix Table 1 confirms that this relationship holds for the time-varying teacher VAM we estimate; a one standard deviation in teacher VAM is associated with a 0.380 standard deviation increase in student test score gains. In Column 2, we consider the correlation between our estimated teacher VAM and a *predicted* student test score gain. The prediction is based on a vector of student characteristics: race, gender and 7^{th} grade language and mathematics scores. Given these characteristics are predetermined by the time the student is assigned to a given EOC teacher, any gains predicted only by these characteristics should be orthogonal to the measured quality of this teacher. If not, this would suggest that there may be sorting of students to teachers based on observable characteristics. The coefficient of 0.008 in Column 2 of Appendix Table 1 indicates that a one standard deviation increase in teacher VAM is associated with less than one percent of a standard deviation increase in predicted test score gains. The considerable attenuation in the coefficient from Column 1 to 2 (it falls by 98 percent) mitigates concerns about selection on observables.

The second test investigates selection on unobservables. This test involves aggregating the estimated time-invariant teacher VAMs and student test score gains to the school-year level and showing that changes in the school-year mean of our estimated teacher value-added affect

the school-year mean of student test score gains. The idea behind this test is that selection among students within a cohort at a given school will be eliminated by aggregating to the school-year level. To be clear, if estimated teacher VAMs only capture student sorting, then changes in the school-year mean of estimated teacher value-added due to the arrival or departure of a teacher should have no effect on mean student outcomes. If not, and the VAMs actually measure teacher quality, then they should be correlated. We show in Appendix Table 2 that both the estimated VAM we actually use (Column 2), and a VAM estimated using a simpler approach in which we do not include class composition and track fixed effects (Column 1) pass this test.

The above two tests provide confidence that the VAMs we estimate and use to measure teacher performance are accurate reflections of teacher quality (in the dimension of student scores on standardized tests).

Appendix C: Tables

Appendix Table 1. Validity of teacher value-added measure: test 1

	(1) Actual gain in student EOC test score	(2) Gain in student EOC test score predicted by student characteristics
Teacher VAM	0.380***	0.008***
	(0.002)	(0.001)
Observations	556,626	556,626
R-squared	0.495	0.972

Robust standard errors (clustered at school district level) in parentheses
*** p<0.01, ** p<0.05, * p<0.1

Appendix Table 2. Validity of teacher value-added measure: test 2

	(1)	(2)
	School-year mean student EOC test score gain	
School-year mean teacher VAM (time invariant, simpler)	0.224***	
	(0.004)	
School-year mean teacher VAM (time invariant, fully specified)		0.201***
		(0.008)
Observations	36,913	36,913
R-squared	0.812	0.795

Robust standard errors (clustered at school district level) in parentheses
*** p<0.01, ** p<0.05, * p<0.1

Appendix Table 3. Effect of performance pay separately by gender and with full gender interaction

	(1)	(2)	(3)
	Male teachers	Female teachers	Full interaction
Treated	0.118	-0.201***	0.115
	(0.124)	(0.044)	(0.121)
Treated X Female			-0.314***
			(0.107)
Course-by-year FEs	X	X	X
Teacher-by-school FEs	X	X	X
Class composition controls	X	X	X
Teacher peer quality controls	X	X	X
Observations	3,127	8,738	11,865
R-squared	0.740	0.681	0.701

Table notes: This table tests the robustness of our main result (Main text, Table 4, Column 5). We repeat our main specification, but run the specification on a subsample of only male teachers (Column 2) or only female teachers (Column 2). Finally, in Column 3, we modify our main specification to interact all variables with female, not only the treatment indicator.

Robust standard errors (clustered at school district and teacher level) in parentheses

*** p<0.01, ** p<0.05, * p<0.1

Appendix Table 4. Heterogeneity in effect of teacher performance pay by size of potential bonus

	(1)	(2)	(3)	(4)
	Teacher VAM			
Treated X	-0.078***	-0.002		
Minimum bonus ($1k)	(0.020)	(0.031)		
Treated X Female X		-0.098***		
Minimum bonus ($1k)		(0.018)		
Treated X			-0.026***	-0.001
Maximum bonus ($1k)			(0.006)	(0.009)
Treated X Female X				-0.034***
Maximum bonus ($1k)				(0.006)
Course-by-year FEs	X	X	X	X
Teacher-by-school FEs	X	X	X	X
Class composition controls	X	X	X	X
Teacher peer quality controls	X	X	X	X
Observations	11,865	11,865	11,865	11,865
R-squared	0.696	0.696	0.696	0.696

Table notes: All specifications in this table are at the teacher-by-year level. We test how the impact of treatment varies with the size of the bonus. Columns 1 and 2 test how treatment varies with the size of the smallest possible bonus within the district (typically awarded for reaching the lowest threshold of value-added). Columns 3 and 4 test how treatment varies with the size of the largest possible bonus (awarded for reaching the highest threshold of value-added). Otherwise, the specifications are the same as our main specifications (Main Text, Table 4, Column 5).

Robust standard errors (clustered at school district and teacher level) in parentheses

*** p<0.01, ** p<0.05, * p<0.1

Appendix Table 5. Effect of performance pay with alternate control for previous student achievement

	(1)	(2)	(3)	(4)
	Student EOC test score			
Panel A: Overall impact of performance pay				
Treated	-0.032***	-0.027**	-0.028**	-0.030***
	(0.011)	(0.011)	(0.011)	(0.011)
Panel B: Impact of performance pay, allowing for gender difference				
Treated	-0.006	-0.000	-0.010	-0.002
	(0.015)	(0.016)	(0.014)	(0.017)
Treated X Female	-0.033***	-0.033***	-0.023**	-0.035***
	(0.011)	(0.012)	(0.009)	(0.012)
Course-by-year Fes	X	X	X	X
Teacher-by-school Fes	X	X	X	X
Control for previous achievement:				
Linear predicted score (main specification)	X			
Predicted score deciles		X		
Raw grade 8 math and reading scores			X	
Raw grades 6-8 raw math and reading scores				X
Observations	569,590	569,590	569,590	459,561

Table notes: All specifications in this table are at the student-by-course level and take a student's standard-normalized end-of-course test score as the outcome variable. All specifications are modifications of the main student-level specification (Main Text, Table 7, Column 4), where the only change is the control for students' prior achievement. In the main text, we include a control for predicted score. That specification is replicated in Column 1 of this table. In Columns 2 to 4 we employ alternate controls. The number of observations in Column 4 is smaller because not all students were present in the data in 6^{th} or 7^{th} grade, so their 6^{th} or 7^{th} grade test scores are missing.

Robust standard errors (clustered at school district and teacher level) in parentheses

*** p<0.01, ** p<0.05, * p<0.1