Study Context & Purpose

Prior to the 2013-14 school year a Tennessee policy went into effect such that a teacher's level of effectiveness (LOE) and certification status determined their minimum number of annual observations (Tennessee Board of Education, 2013). LOE is a composite measure of teacher effectiveness: 50% of a tested teacher's LOE is based on summative observational ratings, 35% value-added measure (VAM), and 15% a school wide achievement score. LOE ranges from 100 – 500. Certification status is essentially based on years of experience. Teachers have Professional or Apprentice status.

If a Professional teacher's LOE was within the range [100, 200), [200,425), or [425, 500] they respectively needed a minimum of four, two, or one formal observation the following year (see Figure 1). If an Apprentice teacher's LOE was within these same ranges they respectively received four, four, or one observation. After each formal observation, observers are to provide feedback to the observed teacher regarding their performance so that teachers may use that information to improve their performance (Tennessee Board of Education, 2013). The stated primary goal of the Tennessee teacher evaluation system is to improve teacher performance (Tennessee Department of Education, n.d.).

The purpose of this study is to identify the effect of increasing the number of observations a teacher receives on student achievement.

Sample

Data come from 2011-12 through 2014-15 TDOE administrative datasets. Administrative data contain measures of student and teacher performance and demographics. The outcomes of interest are teacher-level grade/ subject standardized math and reading achievement scores from students in grades 4 – 12. The unit of analysis is at the teacher-year level (i.e. teacher *i* in year *t*). Outcomes are aggregated to avoid clustering problems with standard errors. Table 1 presents descriptive statistics for the analytical sample.

Research Design

Treatment is the receipt of an additional annual observation (e.g. receiving a total of three annual observations instead of two). Treating variation in the number of formal observations a teacher receives as exogenous is problematic. For example, school administrators may observe less effective teachers more because they want to closely monitor their teaching. To overcome this endogeneity problem, I identify exogenous variation in the number of observations a teacher receives using regression discontinuity (RD). The location of teachers whose continuous LOE scores place them just on either side of the LOE=200 or LOE=425 threshold is treated as local randomization.

Specifically, I use crossing the 200 or 425 thresholds as an instrument to identify policy-induced exogenous changes in observations. Crossing the 200 or 425 threshold induces a change in policy-induced observations. To gain statistical power I re-center continuous LOE in the range [300, 500] at 425, and LOE in [100, 300) at 200, because crossing either threshold induces treatment. Thus, 0 on the re-centered, "overlapped" running variable represents either 200 or 425, and a value of x means the teacher is x units away from either 200 or 425 on the original, uncentered LOE. Crossing the overlapped threshold at 0 induces a change in the state assignment of observations. I explore heterogeneous effects at the 200 and 425 thresholds below. I instrument for the number of observations received using 2SLS because there is substantial non-compliance with the state-assigned observation schedule. The first stage models four

discontinuities: a discontinuity at the original 200 and 425 thresholds for Professional teachers, and two corresponding discontinuities for Apprentice teachers. The 2SLS includes a year fixed effect accounting for changes in the observation system over time, and controls for teacher experience, gender, level of education, and race. There are also controls for the proportion of students taught with respect to: gender, race/ ethnicity, and FRPL, ESL and immigrant status (e.g. the proportion of female students taught out of all students taught could be 0.49). There are also controls for the school mean, standard deviation, and skewness of teacher LOE because these variables may have an independent influence on aggregated student achievement. Finally, each model includes a fourth-order polynomial of teacher-year-level mean prior achievement for the incoming students (i.e. for students taught in year *t*, this controls for the achievement of these same students from year *t* - 1). I interpret the estimate of interest as the change in average standardized reading/ math scores due to the receipt of the marginal annual observation.

Findings

There is no statistical evidence of manipulation at either threshold. Furthermore, manipulation is infeasible. The only person with enough power to manipulate the running variable is the observer, and they would need to anticipate a teacher's VAM and achievement scores in their LOE calculation for successful manipulation.

Separate baseline equivalence tests were conducted at the original 200 and 425 thresholds. These tests found threating imbalances at the 200 threshold for Professional and Apprentice teachers (not shown), and two potentially threatening imbalance at the 425 threshold for Apprentice teachers (see Table 2). Thus, the preferred model is a 2SLS using only Professional teachers originally located on either side of the 425 threshold.

Tables 3 and 4 show first-stage results and the LATEs of interest. Table 3 displays results for the samples including Apprentice and Professional teachers from both thresholds. Table 4 displays results for Apprentice and Professional teachers surrounding just the 425 threshold, and estimates for only Professional teachers surrounding the 425 threshold. While there is a predicted change of approximately -0.10 standard deviations in aggregated math scores in Table 3, this change becomes statistically significant in Table 4. I conclude an additional annual observation has no effect on either math or reading achievement.

My research does not address how observations could be improved. The source(s) of these unproductive observations could be in the generation of observational scores, delivery/ receipt of post-observation feedback, and/ or improper matching of teachers with professional learning experiences aiming to improve teacher productivity. If these problems could be pinpointed, observational systems could be re-designed so classroom observations would, on average, improve student achievement.

Figure 1 – Stylized Illustration of Discontinuities

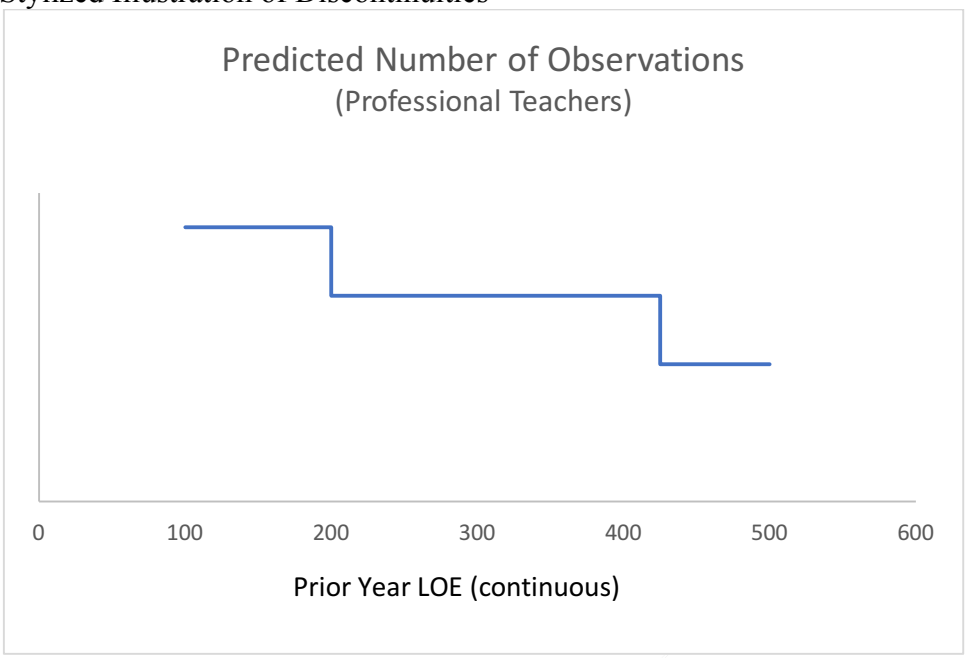

Table 1: Sample Descriptives. DV=Agg Math and RLA Teachers

	Math Mean	Math SD	RLA Mean	RLA SD
Aggregated Math Scores	-0.08	(0.77)	.	.
Annual Change in Aggregated Math Scores	0.10	(0.51)	.	.
Aggregated RLA Scores	.	.	-0.07	(0.76)
Annual Change in Aggregated RLA Scores	.	.	0.10	(0.46)
Female	0.83	(0.38)	0.90	(0.3)
BA+	0.57	(0.49)	0.61	(0.49)
Years Experience	11.47	(8.69)	12.10	(9.07)
Black_White	0.05	(0.22)	0.06	(0.23)
% Sample from Teachers Surrounding LOE 200	6.2	.	6.9	.
% Sample from Apprentice Teachers	15.8	.	14.3	.
Proportion of Students Taught with Characteristics				
Female	0.49	(0.12)	0.48	(0.13)
Black	0.14	(0.19)	0.14	(0.19)
White	0.77	(0.24)	0.77	(0.24)
Asian	0.02	(0.04)	0.02	(0.04)
Hispanic	0.07	(0.1)	0.07	(0.1)
FRPL	0.56	(0.25)	0.54	(0.25)
ESL	0.08	(0.12)	0.07	(0.13)
Immigrant	0.02	(0.05)	0.01	(0.06)

Note: Sample descriptives use data from the analytical sample associated with the largest bandwidth of 40 surrounding the "overlapped LOE." Aggregated scores are teacher-level means of grade/ subject standardized achievement scores. The sample includes students in grades 4 through 12, with two caveats. First, this sample does not include any students if a single teacher did not claim 100% of their math or RLA instructional time. Second, high school students who took their end-of-course assessment in the fall are not included. Standard deviations in parentheses. An annual change in aggregated scores uses achievement data linked to students taught by a teacher in year *t*: it is the difference in aggregated scores received by these students in year *t* and the scores these students received in year $t - 1$. BA+ is an indicator signaling if teacher has earned more than a BA/ BS degree. Black_White is an indicator signaling if teacher is black instead of white. Nearly all TN teachers are black or white. Proportions convey the proportion of students a teacher taught with a given characteristic.

Table 2: Covariate Balance Tests at LOE 425 Threshold. DV=Agg Math and RLA Teachers

	Aggregated Math			Aggregated RLA		
Covariate	$w = 20$	$w = 30$	$w = 40$	$w = 20$	$w = 30$	$w = 40$
Experience: App	-0.10	0.11	-0.02	0.41	0.62	-0.02
	[0.501]	[0.417]	[0.341]	[0.501]	[0.467]	[0.343]
Experience: Prof	-1.76*	0.01	0.75	-2.53**	-0.46	0.06
	[0.873]	[0.715]	[0.617]	[0.961]	[0.775]	[0.657]
Censored Exp: App	-0.02	-0.04	-0.13	0.29	0.20	0.02
	[0.277]	[0.234]	[0.207]	[0.316]	[0.268]	[0.232]
Censored Exp: Prof	-0.33	0.07	0.17	-0.21	0.13	0.25
	[0.258]	[0.211]	[0.181]	[0.252]	[0.203]	[0.175]
Female: App	0.20*	0.19*	0.09	0.02	0.01	0.02
	[0.095]	[0.080]	[0.069]	[0.085]	[0.075]	[0.064]
Female: Prof	-0.06	-0.04	-0.02	-0.03	-0.03	-0.01
	[0.042]	[0.034]	[0.029]	[0.031]	[0.025]	[0.022]
BA+: App	-0.14	-0.09	-0.02	-0.12	-0.05	0.03
	[0.117]	[0.096]	[0.083]	[0.129]	[0.105]	[0.088]
BA+: Prof	0.04	0.02	0.02	0.06	0.02	0.01
	[0.051]	[0.042]	[0.036]	[0.049]	[0.040]	[0.035]
Black: App	-0.10	-0.09	-0.01	-0.13*	-0.05	-0.03
	[0.060]	[0.050]	[0.044]	[0.062]	[0.056]	[0.050]
Black : Prof	-0.04	-0.02	-0.01	-0.02	0.01	0.00
	[0.025]	[0.021]	[0.018]	[0.025]	[0.020]	[0.018]
Lagged Summative Observational Scores: App	-0.02	0.04	0.02	-0.01	0.10	0.06
	[0.071]	[0.061]	[0.053]	[0.104]	[0.081]	[0.068]
Lagged Summative Observational Scores: Prof	-0.05	-0.01	-0.02	-0.13**	-0.09*	-0.08*
	[0.042]	[0.035]	[0.030]	[0.045]	[0.037]	[0.031]
Lagged VAM: App	-0.42	-1.89	-2.25*	1.51	-0.03	-0.79

	[1.428]	[1.235]	[1.065]	[1.151]	[0.840]	[0.673]
Lagged VAM: Prof	0.27	0.29	0.36	0.34	0.31	0.36
	[0.616]	[0.498]	[0.429]	[0.328]	[0.266]	[0.228]
Prior Agg Student Score: App	-0.00	0.10	0.03	0.04	0.13	0.11
	[0.230]	[0.160]	[0.134]	[0.266]	[0.194]	[0.159]
Prior Agg Student Score: Prof	0.00	0.01	-0.00	-0.03	0.00	0.02
	[0.083]	[0.067]	[0.056]	[0.073]	[0.058]	[0.049]
Proportion of Students Taught with Characteristics						
Female : App	0.04	0.01	0.02	-0.01	0.02	0.01
	[0.040]	[0.031]	[0.026]	[0.045]	[0.034]	[0.029]
Female : Prof	-0.02	-0.00	-0.00	-0.01	-0.00	-0.01
	[0.013]	[0.011]	[0.009]	[0.015]	[0.012]	[0.010]
Black: App	-0.00	0.00	0.00	0.00	0.00	0.00
	[0.005]	[0.004]	[0.003]	[0.005]	[0.004]	[0.003]
Black: Prof	0.00	0.00	0.00	0.00	0.00	-0.00
	[0.001]	[0.001]	[0.001]	[0.002]	[0.001]	[0.001]
White: App	-0.00	0.00	0.00	0.00	0.00	0.00
	[0.005]	[0.004]	[0.003]	[0.005]	[0.004]	[0.003]
White: Prof	0.00	0.00	0.00	0.00	0.00	-0.00
	[0.001]	[0.001]	[0.001]	[0.002]	[0.001]	[0.001]
Asian: App	0.00	0.00	0.00	0.00	0.00	0.00
	[0.004]	[0.003]	[0.003]	[0.005]	[0.004]	[0.003]
Asian: Prof	0.00	0.00	0.00	0.00	0.00	-0.00
	[0.001]	[0.001]	[0.001]	[0.001]	[0.001]	[0.001]
Hispanic: App	-0.00	0.00	0.00	0.00	0.00	0.00
	[0.005]	[0.004]	[0.003]	[0.005]	[0.004]	[0.003]
Hispanic: Prof	0.00	0.00	0.00	0.00	0.00	0.00
	[0.001]	[0.001]	[0.001]	[0.001]	[0.002]	[0.001]

FRPL: App	0.12*	0.08*	0.05	0.11	0.08	0.05
	[0.051]	[0.041]	[0.035]	[0.061]	[0.047]	[0.038]
FRPL: Prof	0.02	0.01	-0.00	0.02	0.01	-0.00
	[0.021]	[0.017]	[0.015]	[0.020]	[0.016]	[0.013]
ESL: App	-0.03	-0.02	-0.02	-0.01	0.01	0.00
	[0.028]	[0.021]	[0.016]	[0.016]	[0.013]	[0.011]
ESL: Prof	-0.01	-0.01*	-0.01	-0.01	-0.01*	-0.01*
	[0.006]	[0.004]	[0.004]	[0.006]	[0.005]	[0.004]
Immigrant: App	0.02	0.01	0.01	0.01	0.01	0.00
	[0.019]	[0.015]	[0.012]	[0.021]	[0.018]	[0.015]
Immigrant: Prof	0.00	0.00	0.00	0.00	0.00	0.00
	[0.004]	[0.004]	[0.003]	[0.004]	[0.003]	[0.003]

Note: Point estimates for discontinuities at threshold 425 are displayed. Estimates represent the total predicted change in the outcome for either Professional or Apprentice teachers; estimates are not interactions. Standard errors in brackets. Standard errors clustered at the teacher-level. OLS estimator employed to estimate all coefficients. BA+ is a binary variable indicating if a teacher reported having a degree higher than a BA/ BS. Black is an indicator signaling whether the teacher reported their ethnicity/ race as Black or White. Censored experience is years of experience censored after ten years so all teachers with more than ten years of experience are assigned a value of ten. * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$.

Table 3: Overlapped RDD Main Results. DV=Agg Math and RLA Teachers

	DV = Agg Math			DV = Agg RLA		
	$w = 20$	$w = 30$	$w = 40$	$w = 20$	$w = 30$	$w = 40$
2nd Stage: Number of Annual Observations	-0.11*	-0.10*	-0.08*	-0.04	-0.03	-0.05
	[0.049]	[0.042]	[0.035]	[0.043]	[0.032]	[0.029]
1st Stage:						
App Below LOE 200	1.26	1.15	1.70***	1.34***	1.01***	1.08***
	[0.679]	[0.0595]	[0.486]	[0.285]	[0.227]	[0.197]
Prof Below LOE 200	2.28***	2.27***	2.17***	2.06***	2.57***	2.63***
	[0.257]	[0.197]	[0.201]	[0.473]	[0.454]	[0.457]
App Below LOE 425	1.31***	1.30***	1.16***	0.97*	1.28***	1.17***
	[0.360]	[0.279]	[0.239]	[0.394]	[0.302]	[0.259]
Prof Below LOE 425	0.50***	0.50***	0.57***	0.61***	0.64***	0.64***
	[0.101]	[0.081]	[0.071]	[0.101]	[0.080]	[0.068]
Prof	-0.28	-0.15	-0.17	-0.24	-0.24	-0.15
	[0.219]	[0.175]	[0.149]	[0.296]	[0.233]	[0.202]
Intercept	4.31***	4.33***	3.79***	2.67**	2.73***	2.79***
	[0.920]	[0.081]	[0.714]	[0.973]	[0.564]	[0.546]
N(Teachers-Year)	3348	5205	7197	3143	4906	6783

Note: Teacher-clustered standard errors are in brackets. All models include a fourth order polynomial of the teacher-level mean of prior achievement scores for students taught in year *t* (e.g. the 2011-12 scores of students taught in 2012-13), proportion of students taught holding various characteristics, teacher demographics including certification status, controls for the distribution of teacher effectiveness at the school level, a second order polynomial of overlapped LOE interacted with teacher certification status, and year fixed effects. Each 1st stage estimate represents the total effect of crossing a threshold for each teacher group, none of the 1st stage estimate are interactions. * ($p < 0.05$), ** ($p < 0.01$), *** ($p < 0.001$)

Table 4: Exploring Heterogeneity at Thresholds. DV=Agg Math or RLA

	DV = Agg Math			DV = Agg RLA		
	$w = 20$	$w = 30$	$w = 40$	$w = 20$	$w = 30$	$w = 40$
Teachers Surrounding LOE 425 Only						
2nd Stage: Number of Annual Observations	-0.13	-0.09	-0.08	-0.07	-0.04	-0.06
	[0.088]	[0.079]	[0.067]	[0.083]	[0.055]	[0.048]
N (Tch-Yrs)	3348	5205	7197	2975	4625	6315
Professional Teachers Surrounding LOE 425 Only						
2nd Stage: Number of Annual Observations	-0.28	-0.17	-0.07	-0.07	-0.03	-0.02
	[0.159]	[0.117]	[0.085]	[0.081]	[0.060]	[0.052]
N (Tch-Yrs)	3175	4909	6746	2975	4625	6315

Note: Teacher-clustered standard errors are in brackets. All models include a fourth order polynomial of the teacher-level mean of prior achievement scores for students taught in year *t* (e.g. the 2011-12 scores of students taught in 2012-13), proportion of students taught holding various characteristics, teacher demographics including certification status, controls for the distribution of teacher effectiveness at the school level, a second order polynomial of overlapped LOE interacted with teacher certification status, and year fixed effects. Each 1st stage estimate represents the total effect of crossing a threshold for each teacher group, none of the 1st stage estimate are interactions. * ($p < 0.05$), ** ($p < 0.01$), *** ($p < 0.001$)

Hill, H. C., & Grossman, P. (2013). Learning from Teacher Observations: Challenges and Opportunities Posed by New Teacher Evaluation Systems. *Harvard Educational Review*, *83*(2), 371–385. https://doi.org/10.1017/CBO9781107415324.004

Kluger, A. N., & DeNisi, A. (1996). The Effects Of Feedback Interventions On Performance: A Historical Review, A Meta-Analysis, And A Preliminary Feedback Intervention Theory. *Psychological Bulletin*, *119*(2), 254–284. https://doi.org/10.1037/0033-2909.119.2.254

Neumerski, C. M., Grissom, J. A., Goldring, E., Cannata, M., Drake, T. A., Rubin, M., & Schuermann, P. (2014). Inside Teacher Evaluation Systems: Shifting the Role of Principal as Instructional Leader. In *Association of Education Finance and Policy* (pp. 1–32). San Antonio, TX.

Tennessee Board of Education. Teacher and Principal Evaluation Policy (2013). Nashville, TN.

Tennessee Department of Education. (n.d.). Evaluation | TEAM-TN. Retrieved September 22, 2016, from http://team-tn.org/evaluation/

Tuytens, M., & Devos, G. (2011). Stimulating professional learning through teacher evaluation: An impossible task for the school leader? *Teaching and Teacher Education*, *27*(5), 891–899. https://doi.org/10.1016/j.tate.2011.02.004