TN EDUCATION RESEARCH ALLIANCE

Working Paper
December 14, 2023

Policy-Assigned Teacher Observations, Their Implementation, and Student Discipline Outcomes: Main, Mediated, and Moderated Relationships

Seth B. Hunter | *George Mason University*
Adam Kho | *University of Southern California*
Katherine M. Bowser | *George Mason University*

Policy-Assigned Teacher Observations, Their Implementation, and Student Discipline Outcomes: Main, Mediated, and Moderated Relationships

Seth B Hunter, Corresponding Author
ORCiD 0000-0002-3051-872X
George Mason University, Tennessee Education Research Alliance
Assistant Professor of Education Leadership
4400 University Dr, MS4C2
Fairfax, VA 22030
shunte@gmu.edu

Adam Kho
ORCiD 0000-0002-6957-3427
University of Southern California, Tennessee Education Research Alliance
Assistant Professor of Education Policy and Leadership
3470 Trousdale Parkway, WPH 901E
Los Angeles, CA 90017
akho@usc.edu

Katherine M Bowser
George Mason University
4400 University Dr, MS4C2
Fairfax, VA 22030
kbowser@gmu.edu

Data availability statement: The datasets generated during and/or analyzed during the current study are not publicly available because the authors do not have permission to grant access. Those interested in gaining access may submit a research proposal to the Tennessee Education Research Alliance.

The authors thank the Tennessee Department of Education, Tennessee Education Research Alliance, and participants from the annual Association for Education Finance and Policy conferences for helpful feedback.

Abstract

This study is the first to estimate main, mediated, and moderated relationships between policy-assigned observations and various student discipline outcomes (SDOs). We also examine the relationships between SDOs and observations conducted. The data suggest that the percentage of students who receive at least one SDO decreases as policy assigns schools an additional 25 observations and that in-school suspension reductions drive this relationship. Improvements in teachers' classroom management skills mediate some SDO reductions. We find that several relationships substantively depend on the degree of prior-year SDOs and the average teacher's years of experience. While policy may reduce SDOs, the data also suggest that schools implement observations in ways that may increase them. We discuss implications for policy, practice, and research.

Keywords: Student discipline; teacher evaluation; classroom observations; education policy; regression; implementation

Introduction

As of 2023, most state education agencies in the United States have implemented "next-generation" teacher evaluation systems to improve student achievement (American Institutes for Research, 2016; Bleiberg et al., 2021; Steinberg & Sartain, 2015). Next-generation systems differ from past systems in at least two important ways: the widespread use of revised standards-based observation rubrics (e.g., Framework for Teaching) and an expectation that the typical teacher receives more frequent observations (National Council on Teacher Quality, 2019; Steinberg & Sartain, 2015). Theoretically, next-generation observations can improve student achievement because standards-based rubrics describe teaching practices that are positively linked to student achievement, and more rubric-based feedback from observations can foster effective standards-based teaching (Donaldson & Papay, 2014; Papay, 2012).

There are several reasons why policymakers and researchers should be concerned about the effects of next-generation observations. First, research suggests that observations are the most expensive component of next-generation teacher evaluation systems (Stecher et al., 2016); such costly reforms should yield substantial benefits. Second, school administrators, the typical observer (Johnson & Fiarman, 2012), report that next-generation observations are burdensome and time-consuming (Neumerski, 2013; Rigby, 2015). To cope with these demands, observers implement shorter observations and feedback conferences, which may limit the effectiveness of next-generation observations (Donaldson & Woulfin, 2018; Hunter & Rodriguez, 2021; Kraft & Gilmour, 2016). Finally, if next-generation observations improve teacher effectiveness, the benefits to students could be substantial; prior work finds that students experience better short- and long-run academic and non-academic outcomes when taught by teachers who improve achievement (Chetty et al., 2014; Jackson, 2018; Liu & Loeb, 2021).

However, research on the effects of next-generation evaluation systems on student achievement, or teacher effectiveness in terms of student achievement, has produced mixed results. The evidence suggests that the introduction of well-implemented citywide next-generation observations improve student achievement in math or reading, but we are unaware of any study finding evidence of improving both (Steinberg & Sartain, 2015; Taylor & Tyler, 2012). New research suggests that the net effects of next-generation observation policies in relatively low-stakes systems assigning all teachers at least one observation each year does not improve student achievement (de Barros, 2019; Hunter, 2019; Hunter & Kho, 2023), but the threat of an impending announced observation in high-stakes systems improves teacher performance (Phipps & Wiseman, 2021).

Although we are beginning to understand the effects of next-generation observations on student achievement outcomes, we know very little about the effects on non-academic student outcomes, such as discipline. Yet, there are reasons to be concerned about these effects as recent research underscores the importance of reducing student misbehavior. Student exposure to more disruptive classroom peers negatively affects longer-term academic achievement, post-secondary earnings, and in some cases, reduces the likelihood of college enrollment and completion (Carrell et al., 2018). Other research finds that teachers who are more effective at reducing suspensions and other misbehaviors improve longer-term academic outcomes, reduce future suspensions, and increase the likelihood of high school graduation (Jackson, 2018). Furthermore, teacher effectiveness measured in terms of student achievement is only weakly correlated with effectiveness measured in terms of non-academic outcomes (Jackson, 2018; Liu & Loeb, 2021), which suggests that what we know about the effects of next-generation observations on student achievement may not apply to effects on student disciplinary outcomes (SDOs) .

This paper extends our understanding of teacher evaluation and teacher labor markets by investigating relationships between observations and SDOs. We frame our analyses using the school disciplinary process conceptual framework from Rodriguez and Welsh ("RW framework") (2022), which describes the various factors affecting SDOs. We examine four school-year outcomes: the percentage of students who (1) receive at least one in-school suspension (ISS), (2) receive at least one out-of-school suspension (OSS), (3) are expelled (EX) at least once, and (4) receive at least one ISS, OSS, or EX. The focal analyses investigate relationships between SDOs and the total number of observations assigned to all teachers by state policy, and potential mediators and moderators. Secondary analyses estimate relationships between the number of observations received and SDOs.

This study makes four contributions to our understanding of teacher evaluation and student discipline. It is the first to estimate average intent-to-treat (ITT) relationships between policy-assigned observations and SDOs, which are plausibly the most relevant estimates for education policymakers. Policymakers cannot control the extent to which schools implement evaluation policy with fidelity; policymakers can only prescribe what should happen. In our case, policymakers can only tell schools how many observations to conduct, they cannot ensure that schools conduct the number of observations assigned, nor can they ensure that schools implement observational or student disciplinary processes with fidelity. Second, we examine the extent to which ITT relationships depend on observable school characteristics. Third, we estimate the relationship between the number of observations schools conduct and SDOs. Finally, we extend the RW framework and prior work concerning teacher evaluation and SDOs by uniting these two literatures.

Conceptual Framework and Related Literature

We draw heavily from Rodriguez and Welsh's framework (RW) to describe the factors influencing SDOs (2022), then expand RW's framework by incorporating teacher evaluation and development literature to argue why more classroom observations might affect SDOs. In broad terms, RW's framework describes several factors conceptually related to how teachers respond to perceived student misbehavior, whether and why a student receives a formal referral or not, and whether and why the student receives an SDO, which RW defines as an “exclusionary discipline” outcome (Rodriguez & Welsh, 2022). While RW also elucidates why and how students receive non-exclusionary outcomes (e.g., parent conferences, formal warnings), we focus on exclusionary SDOs only because our data do not include non-exclusionary outcomes (henceforth, our use of SDOs refers to exclusionary outcomes only).

Factors Affecting SDOs

Several within-school factors may affect how teachers respond to perceived student misbehaviors. For example, teachers may be less tolerant of perceived misbehaviors from students with a history of frequent office referrals (Rodriguez & Welsh, 2022). Classroom management and teaching effectiveness might also influence how teachers respond to perceived misbehaviors. Skilled classroom managers may respond to misbehavior via classroom management (i.e., without writing a formal referral or sending a student to the administrators office), while less-skilled classroom managers may be more likely to request administrative management (Rodriguez & Welsh, 2022). Furthermore, more effective teachers, broadly defined, may better manage perceived student misbehavior (Jackson, 2018).

Teacher responses to perceived student misbehavior affect subsequent steps in school disciplinary processes (Rodriguez & Welsh, 2022). Classroom management responses alone cannot lead to SDOs because they stop the student disciplinary process from advancing further toward *exclusionary* discipline consequences, which includes in-school suspensions (ISS), out-of-school suspensions (OSS), and expulsions (EX). However, teachers who need additional support may informally or formally (i.e., submit office referrals) request administrative help, or administrators may overtake the student's behavior management. Administrators may only issue SDOs if the student receives a formal office referral; however, administrators may decide to assign students non-exclusionary consequences (e.g., parent conference) even if the student has not received a formal referral. While office referrals precede SDOs, we do not distinguish between the two stages because the data do not include referrals without SDO - the only referrals observed are those leading to SDOs.

The RW framework (2022) asserts that several factors affect whether students receive an SDO. In particular, student, teacher, and administrator demographics can influence the likelihood that a student receives a referral or SDO (Rodriguez & Welsh, 2022; Skiba et al., 2014; Welsh & Little, 2018). In broad terms, the legal ramifications associated with disciplining SPED students (Osborne, Jr., 2001), teacher and administrator biases arising from their views regarding specific student demographics (e.g., student economic disadvantage, race, gender), and the racial or gendered biases that affect teacher or administrator views independent of student demographics, may explain why these demographics affect student referrals/ SDOs (Rodriguez & Welsh, 2022; Skiba et al., 2014; Welsh & Little, 2018).

Although the RW framework (2022) does not explicitly state that teacher and administrator experience and effectiveness affect the disciplinary process, it and prior work imply as much. RW argues that teacher instructional and classroom management skills affect how teachers respond to perceived misbehavior (2022). We extend this argument to include composite measures of teacher effectiveness based on observational ratings of teacher instruction and classroom management and teacher effects on student academic and non-academic outcomes. Similarly, RW (2022) implies that principal effectiveness may affect whether a student receives an SDO. We assume that principals with higher composite effectiveness scores based on student academic and non-academic outcomes, faculty input, and portfolio performance (i.e., like the principal effectiveness measures from our study setting) can manage student misbehavior without resorting to exclusionary discipline outcomes better than less effective principals. Notably, how principals manage student discipline consequences partially determines their portfolio performance measure, a component of the principal effectiveness measure in the study setting (for details, see Grissom et al., 2018). Additionally, we assume teachers' and administrators' years of experience affect school discipline processes. Although we are unaware of research documenting returns to teacher or administrator experience regarding SDOs, we assume that, over time, teachers and administrators eschew ineffective responses to perceived student misbehavior, consistent with the returns to teaching experience regarding student achievement (Kraft et al., 2020; Papay & Kraft, 2013).

Finally, RW (2022) argue that several school-level characteristics affect student SDOs. For example, schoolwide policies, experiences with behavior management programs, and neighborhood settings may affect teacher-student relational dynamics, schoolwide discipline reporting practices, perceptions of infraction severity, and principal discretion in issuing SDOs. These and other factors affect how teachers, administrators, or teachers-and-administrators manage misbehavior and whether administrator-managed misbehaviors result in an SDO.

Classroom Observations, SDOs, and Mechanisms

Liebowitz and colleagues (2022) estimate the causal effect of Race-to-the-Top (RTTT) teacher evaluation reforms on office referrals relative to pre-RTTT systems. In their study, they implicitly assumed that classroom observations and other features of Race-to-the-Top (RTTT) era teacher evaluation reforms could affect schools' disciplinary processes. The authors concluded that evaluation reforms did not affect office referrals, but did not examine whether observations affect SDOs specifically. Liebowitz and colleagues (2022) estimated black box effects that did not distinguish between different evaluation-system reforms, some of which may have divergent or unintended effects on disciplinary processes. Indeed, the mechanisms by which an individual reform affects SDOs may yield different effects; as we argue below, this may be the case for the widespread reform that increased classroom observations.

We hypothesize that classroom observations might affect SDOs via three mechanisms. First, more observations might reduce SDOs by improving classroom instruction, specifically helping teachers improve classroom management skills such that teachers can better manage perceived student misbehavior in the classroom without requiring the use of office referrals. We refer to this as the *Class Management* mechanism. Standards-based teacher performance rubrics and structured post-observation conferences accompanied the increase in RTTT-era classroom observations (Steinberg & Donaldson, 2016), supporting the proposed relationship between observations and SDOs. Standards-based rubrics, like the ubiquitous Framework for Teaching, mapped different aspects of teaching onto performance levels. For example, teachers in Tennessee (our study setting) were observed using a rubric based on the Framework for Teaching, and their teaching was assessed against standards-based questioning strategies, feedback to students, teacher expectations, and classroom environment, among other standards. Following these assessments, evaluators (typically principals or assistant principals, see Hunter, 2021) facilitated post-observation conferences during which teachers received quantitative and qualitative performance feedback to improve the measured aspects of teaching (Donaldson, 2021; Hunter & Springer, 2022). Because standards-based rubrics include aspects of teaching concerning classroom management, frequent observations and standards-based post-observation performance feedback may improve teacher classroom management and reduce SDOs. However, recent work has cast doubt on the effective implementation of these observational processes (Donaldson & Woulfin, 2018; Hunter, 2022; Hunter & Rodriguez, 2021), which may attenuate the ability of observational processes to reduce SDOs.

Administrator involvement in observational processes accounts for the second and third hypothesized mechanisms by which observations might affect SDOs. More observations cause administrators, who represent the bulk of teacher evaluators (Hunter, 2023; Hunter & Ege, 2021), to be in classrooms with students more often. Frequent administrator presence in classrooms may dissuade some student misbehaviors from manifesting in the first place, as students are aware that administrators ultimately have the power to assign SDOs. Further, administrator presence may strengthen teacher-administrator and student-administrator relationships, which might also reduce SDOs (Rodriguez & Welsh, 2022; Welsh & Little, 2018). Thus, observation-induced administrator visibility may reduce SDOs; we refer to this as the *Administrator Presence* mechanism.

Administrator involvement in observation processes explains the third and final mechanism by which observations affect SDOs; however, unlike the two previous mechanisms, the final hypothesized mechanism may increase SDOs. Administrators report that it is overly burdensome to conduct more observations (Kraft & Gilmour, 2016), and surveys suggest it takes administrators 30 minutes, on average, to conduct each formal observation in the Tennessee context (Hunter, 2020; Hunter & Rodriguez, 2021). The time it takes to conduct observations may crowd out other administrator tasks (Demerouti et al., 2001; Hunter & Rodriguez, 2021), which may pull administrators away from office-based student behavior management. Specifically, administrators may not have the time to counsel students who receive formal office referrals toward non-exclusionary disciplinary consequences (e.g., parent conferences), the enactment of which demands more administrator time (Huang et al., 2020). Consequently, more observations may pull administrators away from some office-based SDO-reducing tasks, which means more observations might increase SDOs; we label this mechanism *Administrator Time.*

SDO Malleability

Although few studies examine multiple SDO types (Skiba et al., 2014; Welsh & Little, 2018), different types of SDOs may offer insights into school disciplinary processes. Prior work recognizes three types of SDOs, all of which follow formal office referrals in school disciplinary processes: in-school-suspensions (ISS), out-of-school suspensions (OSS), and expulsions (EX) (Skiba et al., 2014). Generally, administrators assign referrals associated with less severe misbehaviors to ISS (e.g., foul language, disobedience, disrespect), those associated with more severe behavior to OSS (e.g., violence, property damage), and expel repeated or extremely severe misbehavior (e.g., criminal behavior) (Skiba et al., 2014). To the extent that observations affect SDOs, we hypothesize that ISS are more malleable than OSS, which are more malleable than EX, for three reasons. First, the three mechanisms by which observations might affect SDOs are relatively light-touch interventions. Post-observation feedback-based improvements to classroom management, administrator presence in classrooms, and administrator time on disciplinary tasks may not be intensive enough to change the causes or responses to relatively severe misbehavior. Second, prior work reports that less severe misbehaviors are more sensitive to interventions than more severe misbehaviors (Lacoe & Steinberg, 2019). Finally, because EX and OSS rates are relatively low (see below), there is less variation in these outcomes to affect.

Moderation

We explore potential moderators between more observations and SDOs by prior-year SDOs and the two teacher characteristics that determine the number of policy-assigned observations teachers receive: years of experience and the *de facto* composite measure of teacher effectiveness in the study setting. Prior work suggests that it may be more difficult to lower SDOs in settings with historically high SDOs (Welsh & Little, 2018). We are also interested in moderation by teacher experience and prior-year effectiveness for policy-related reasons. Policy assigns less experienced and less effective teachers in the study setting more observations. If observations based on teacher experience and effectiveness reduce SDOs, reductions may be larger in settings with higher concentrations of teachers possessing these characteristics. Alternatively, if teachers with fewer years of experience or lower levels of prior-year effectiveness have greater difficulty managing student behavior than their colleagues with higher levels of human capital, schools with higher concentrations of less-experienced or less-effective teachers may have greater difficulty managing student behavior.

Study Context

Our study occurs in Tennessee, which adopted the Tennessee Educator Acceleration Model (TEAM) teacher evaluation system in 2012. TEAM policy states that certified observers should conduct formal observations using a standards-based rubric, and that structured post-observation conferences follow every observation (Teacher and Principal Evaluation Policy, 2013). Below, we highlight key components of TEAM.

Standards-Based Rubric

TEAM observations use the “TEAM rubric” (see Online Appendix A), a standards-based rubric resembling Charlotte Danielson’s Framework for Teaching. The rubric measures teaching across Planning, Instruction, and Environment domains; we characterize the Environment domain as measuring Classroom Management. Rubric indicators describe specific aspects of teaching tasks (e.g., Questioning, Classroom Expectations), which are mapped onto three performance levels: Below Expectations (1), At Expectations (3), and Above Expectations (5).

Structured Observations

Statewide Tennessee survey data suggest that the typical Tennessee teacher observation lasts approximately 30 minutes (Hunter, 2020). During post-observation conferences, which should occur one week after an observation, observers provide performance feedback aligned with TEAM rubric standards (Hunter & Springer, 2022).

Observer Certification

The Tennessee Department of Education (TDOE) monitors observer certification through annual exams. Each summer, observers must pass a two-part exam assessing scoring accuracy, conference facilitation, and knowledge of TEAM policy. New observers must attend a multi-day training; during the study period the number of training days was shortened from four to two per summer. Veteran observers could test out of annual training by passing the certification exam. If a veteran observer does not pass the re-certification exam, they must participate in summer training and retake the exam.

Level of Effectiveness

Near the beginning of a school year, teachers receive an individual composite effectiveness score on an integer scale ranging from one to five. However, integer scores are determined by a composite measure that is a function of prior-year observation, growth, and achievement scores. Teachers of tested subjects receive individual Tennessee Value-Added Assessment System (TVAAS) scores, which serve as their growth score. As teachers of untested subjects do not receive individual TVAAS scores, schoolwide TVAAS scores typically serve as their growth score. Achievement measures are determined by grade-, school-, or district-wide student achievement (e.g., ACT scores, high school graduation rates).

Assignment of Observations

Two factors determine the number of observations assigned to teachers: prior-year integer-based effectiveness scores and certification status, which is determined by teaching experience. State policy assigns one observation to teachers with a prior-year effectiveness score of five and four observations to teachers with a prior-year effectiveness score of one. The number of observations assigned to teachers with a prior-year composite score of two through four depends on certification status: early-career teachers (fewer than four years) are assigned four observations, and more experienced teachers two. We argue that the number of observations assigned by Tennessee policy is of broad interest because assignments of one, two, and four observations per teacher-year are the most popular assignments across the United States.[1] While state policy assigns teachers minima, Tennessee observers can and do apply discretion, issuing more or less observations to specific teachers than dictated by policy (Hunter & Ege, 2021).

Data

We use Tennessee administrative data from the 2012-13 through 2017-18 school years. SDO data are at the student-by-disciplinary-consequence (ISS, OSS, EX) level and link SDOs to students and schools; SDO data do not record who wrote the office referral and therefore these data are not linked to teachers. The data also identify whether a student is economically disadvantaged (ED), if English is their second language (ESL), and if the student is in special

[1] We reviewed current observation policies in each state and find that two is the modal number of assigned observations, followed by one observation, then four. Furthermore, these classroom observation frequencies are consistent with research from the mid-2010s (Steinberg & Donaldson, 2016).

education (SPED). Teacher and administrator administrative data include individual effectiveness scores, years of experience, and education level; teacher records also include observation scores regarding classroom management (i.e., the Environment domain). Student, teacher, and administrator administrative data also include gender and race/ ethnicity. We obtain the number of observations assigned to each teacher each year using their years of experience and prior-year individual effectiveness score, the two determinants of policy-assigned observations. We use teacher observation records at the teacher-by-year-by-observation-occurrence level to count the number of observations received by each teacher each year.

We aggregate all these data to the school-by-year level. The focal independent variables of interest are the total number of observations assigned and received for each school-by-year. School-by-year measures of student demographics (i.e., SPED, ESL, ED, gender, race/ ethnicity) represent the proportion of students exhibiting each characteristic (e.g., the proportion of students enrolled in each school-by-year who are Black). We create similar school-by-year proportions for teacher and administrator gender, race/ ethnicity, and education level. We also create school-by-year mean teacher and mean administrator measures regarding years of experience, composite effectiveness scores, and observation scores.

We use four school-by-year SDO outcomes. The first represents the percentage of students in each school-by-year who received any number of SDOs. For example, if there were four students in a school-by-year and two did not receive any SDOs, the third received three ISSs, and the fourth one expulsion, 50% of the students in that school-by-year received any number of SDOs. We construct similar outcomes for ISS-, OSS-, and EX-specific SDOs.

Our mediation analyses use administrative and teacher and administrator survey data. We use the TEAM rubric's average Environment domain observation scores to explore the *Class*

Management mechanism. The school-by-year average teacher's Environment domain scores range from 1.75 to 5, with a mean of 4.23 and a standard deviation of 0.38.

We explore the *Administrator Presence* and *Administrator Time* mechanisms using survey items. Each spring, TDOE administers the Tennessee Educator Survey (TES) to teachers and administrators to gather information on various topics, including teacher and administrator time use and perceptions of student behavior and observational processes. The 2013 and 2014 teacher TES' included the following item: "Teaching observations disrupt my classroom instruction," and the 4-option Likert response was Strongly Disagree (=1), Disagree, Agree, and Strongly Agree (=4). During the 2013 and 2014 TES administrations, 98% of schools provided at least one response to the Administrator Presence survey item, between 1% and 100% of potential teacher respondents in each school-by-year submitted responses, the mean school-by-year response rate is 33%, the mean reverse-coded response is 2.62, and the standard deviation of responses is 0.34. The *Administrator Presence* mechanism implies that evaluators do not disrupt classrooms and that a more frequent evaluator presence in classrooms dissuades student misbehaviors; thus, we reverse-coded the teacher survey item. Ultimately, the *Administrator Presence* hypothesis suggests that more observations will be associated with higher ratings in the reverse-coded teacher survey item regarding classroom disruptions.

Finally, we investigate the *Administrator Time* mechanism using an administrator time-use survey item from the 2015 – 2018 TES administrations. The item stem asked administrators to select the amount or percentage of time they spend on "student discipline issues" during an average week among eight other options, including "administrative duties," "instructional planning with teachers," and "other." The 2015 – 2017 responses to the item stem are None, 1 hour or less, 1 – 3 hours, 3 – 5 hours, 5 – 10 hours, and more than 10 hours. On the 2018 item, administrators identified the percentage of time (0 – 100) they spent on disciplining students and other tasks; percentages listed across the 2018 tasks had to sum to 100.[2] We place responses to the 2015 – 2017 and 2018 on the same scale by finding the total amount of time respondents assigned to time-use tasks on the 2015 – 2017 TES, then find the percentage of total time spent on student discipline issues. We assigned the "None" option zero hours, the "More than 10 hours" option 11 hours, and took the midpoint of other response options (e.g., "1 hour or less" is assigned 0.5 hours). Over the 2015 – 2018 years, 74% of schools provided at least one response to the *Administrator Time* survey item, between 14% and 100% of potential administrator respondents in each school-by-year submitted responses, the mean school-by-year response rate is 75%, the mean response is 6.42%, and the standard deviation is 5.44. Consistent with the hypothesized *Administrator Time* mechanism, we expect the percentage of time the average school-by-year administrator spends on student discipline issues to decrease as observations rise.

Methods

Observation Assignments

We examine the extent to which the number of observations *assigned* is associated with the percent of students who receive at least one: suspension or expulsion, ISS only, OSS only, and EX only. For each outcome, the percentage is equal to the total number of students who received the SDO divided by total enrollment; students receiving multiple SDOs in a year were counted once. We apply ordinary least squares regression with year fixed effects in Equation 1:

$$y_{st} = \delta assigned_{st} + X_{st} + \alpha_t + e_{st} \, , \qquad (1)$$

where y_{st} is one of the four SDOs in school *s* in year *t*. The coefficient of interest, δ, represents the average change in y_{st} that is associated with an increase in $assigned_{st}$, the number of

[2] For example, if there were only two percentage of time on task options and a respondent filled in 60% for one task, the survey would only permit the respondent to list 40% for the other task.

policy-assigned observations in school s in year t.[3] To aid in the interpretation of results, we scale $assigned_{st}$ by 25 such that for every unit increase in $assigned_{st}$ the total school-by-year policy-assigned observations increases by 25. The RW framework implies that we should control for X_{st}, a vector of characteristics in school s in year t, that includes the prior-year SDO of interest; average teacher and administrator prior-year LOE; the proportions of students, teachers, and administrators who are white, black, and female; the proportions of students who are English-language learners, special education, and economically disadvantaged; and the average teacher and administrator years of experience and level of education. Additionally, we control for the number of teachers in each school and the ratio of teachers to evaluators in each school. While we include all covariates in X_{st} to aid in the comparison of observably similar schools, the last two are critical to our research design. Controlling for the total number of teachers in a school is vital because schools that have a similar number of total assigned observations, but drastically different teacher totals may differ in important ways that impact SDOs, such as increased presence of teacher leaders or peer collaboration. We also control for the ratio of evaluators-to-teachers to compare similar schools. As described in our theory of action, frequent administrator presence in classrooms may dissuade student misbehavior (*Administrator Presence*) and conducting observations may pull administrators away from discipline-related responsibilities (*Administrator Time)*, increasing SDOs. The ratio of evaluators-to-teachers within schools likely affects the delegation of evaluation and disciplinary related duties, potentially affecting the *Administrator Presence* and *Administrator Time* mechanisms. Finally,

[3] The relationship between SDOs and observations assigned or received may be nonlinear; specifically, the marginal returns to more observations may diminish. We explore nonlinearities by adding a quadratic observation assigned or received term to equations 1 and 2; we also explore logarithmic relationships by replacing the linear observations assigned or received term in equations 1 and 2 with the natural log of observations assigned or received. The evidence does not suggest statistically or practically meaningful differences between linear and nonlinear specifications. Results available upon request.

we include year fixed effects in Equation 1 to control for secular shocks affecting SDOs, such as new state guidance on disciplinary action. Standard errors are clustered at the school level.

Though Equation 1 compares observably similar schools, δ may be biased by time-invariant unobserved between-school factors correlated with both observations assigned and SDOs. For example, schools with a strong culture of belonging may have less policy-assigned observations and SDOs, positively biasing results. Thus, we apply school fixed effects:

$$y_{st} = \delta assigned_{st} + X_{st} + \gamma_s + \alpha_t + e_{st} \ , \qquad (2)$$

where γ_s represents school fixed effects, all other terms are identical to Equation 1, and standard errors are clustered at the school level.

Mediators. The *Class Management, Administrator Presence,* and *Administrator Time* mechanisms may mediate associations between policy-assigned observations and SDOs. If true, we should detect a relationship between observations and each mediator using Equation 3 where, m_{st} is one of the three mediators and all other terms are identical to Equation 2:

$$m_{st} = \delta assigned_{st} + X_{st} + \gamma_s + \alpha_t + e_{st} \ , \qquad (3)$$

We estimate Equation 4 to examine the extent to which m_{st} explains the total unmediated relationships (δ) from Equations 1 and 2:

$$y_{st} = \delta assigned_{st} + \varphi m_{st} + X_{st} + \gamma_s + \alpha_t + e_{st}, \qquad (4).$$

All other terms are identical to Equation 2. Standard errors are clustered at the school level.

Moderators. We examine moderators via Equation 5:

$$y_{st} = \delta assigned_{st} + \Phi mod_{st} + \pi assigned * mod_{st} + X_{st} + \gamma_s + \alpha_t + e_{st} \ , \qquad (5)$$

where mod_{st} is the prior-year SDO ($y_{s(t-1)}$), average teacher prior-year LOE (scaled by 50s), or average years of teacher experience. Standard errors are clustered at the school level.

Sensitivity Tests. We check the sensitivity of our conclusions about policy-assigned observations in two ways. First, we apply a version of Equation 1 to the sample that includes only those teachers whose prior-year effectiveness score was near the cutoffs that determine the number of observations assigned by policy. Teacher experience and prior-year LOE score discontinuities determine the number of observations assigned by policy; the latter is determined by an underlying continuous LOE function, which educators do not see, ranging from 100 to 500. Crossing from LOE 1 to LOE 2 can assign teachers fewer observations, depending on years of experience, and crossing from LOE 4 to LOE 5 assigns all teachers fewer observations. By restricting comparisons to teachers who fell just to either side of the thresholds, we bolster internal validity. However, this test may attenuate estimates due to measurement error as we cannot link SDOs to teachers. See Online Appendix B for additional details.

Second, we apply a formal "sensitivity test" introduced by Rosenbaum and Rubin (1983) and extended by Cinelli and Hazlett (2020). These scholars remind us that a potentially biasing omitted variable (OV) must correlate with residual outcome and treatment variation (i.e., variation not explained by the model), but such correlations are not necessarily sufficient to undo inferences (e.g., OVs weakly correlated with unexplained treatment and outcome variation). Notably, Cinelli and Hazlett's (2020) test reports the maximum bias multiple, non-linear confounders (i.e., "OV") could introduce to one's inferences. Analysts must explain why the reported confounding conditions are not plausible, ruling out any plausible OVs that could explain the amounts of treatment and outcome residual variation needed to undo inferences. We contextualize what is plausible using the explanatory power of the observed covariates that determine treatment or outcome variation. We benchmark one OV against prior-year SDOs, a powerful predictor of SDOs, and argue that it is implausible that an OV could explain more SDO residual variation than what is explained by prior-year SDOs; we press the limits of plausibility by creating scaled-up OVs explaining two and three times as much variation. We repeat this exercise twice, using OVs benchmarked against average teacher years of experience and prior-year LOE scores, and assume that OVs explaining more variation in the number of policy-assigned observations than its policy determinants are implausible.

Observations Received

Teachers in the average school are assigned a total of 63.25 observations, however teachers in the average school receive a total of 85.81 observations. This difference arises because observations received are affected by two components: observations *assigned* and *discretionary* observations. Discretionary observations are those observations which administrators decide to conduct beyond what policy dictates for a particular teacher. To examine the relationship between observations received and SDOs, we substitute observations assigned with observations received (scaled by 25) in both Equations 1 and 2.

Importantly, discretionary observations may come about for various reasons that bias δ. For example, administrators issue more observations than the policy-assigned number to teachers who teach students that are prone to misbehave for unobserved reasons (i.e., reverse causation), thereby resulting in a positively biased δ. Furthermore, relationships between teachers, administrators, and students may change over time within the same school in ways affecting both observations received and SDOs. Ultimately, time variant selection bias is a concern for observations received because administrators have discretion over them, whereas they do not have discretion over observations assigned.

We also employ two-stage least squares (2SLS) to effectively partition observations received into its two sources, observations assigned and discretionary observations. We do not argue that 2SLS brings us closer to causal inferences; instead, 2SLS allows us to exclude the variation arising from discretionary observations and examine relationships with variation from policy-assigned observations. The first-stage, Equation 6, isolates the portion of observations received that arises from policy assignments by regressing the number of observations *assigned* on the number of observations *received*:

$$received_{st} = \gamma assigned_{st} + X_{st} + \alpha_t + \mu_{st}\ , \qquad (6)$$

Then, in the second stage of our 2SLS model, we regress the number of observations received because of policy assignment, $\widehat{received}_{st}$, on each of the four SDOs in Equation 7:

$$y_{st} = \delta \widehat{received}_{st} + X_{st} + \alpha_t + e_{st}\ , \qquad (7)$$

The remaining terms in equations 6 and 7 are as defined in Equation 1. Standard errors are clustered at the school level. We do not purport causality as μ_{st} in Equation 6 is plausibly correlated with y_{st} in Equation 7, violating 2SLS exclusion criteria. However, given the relationship between observations assigned and observations received, we find the use of 2SLS compelling as it is effectively a sensitivity analysis of our primary model and will allow us to explore discrepancies in the associations between SDOs and observations assigned and observations received.

Findings

Table 1 shows school-by-year characteristics of our sample. 8.45% of students received disciplinary action at least once. ISS' are the most common SDO, with 6.09% of students receiving at least one ISS; OSS' are less frequent (3.77%) and EX' is rare (0.07%). On average, schools had just over 63 observations assigned, however almost 86 observations are received per year, meaning that observers chose to conduct more observations than policy mandated.

Observations Assigned

On average, assigning schools 25 more observations is associated with a reduction in the percent of students receiving any SDO, receiving ISS-only, receiving OSS-only, and assigning observations is not associated with EX-only (Table 2). Column I reports results from Equation 1. There is a statistically significant reduction of 0.30 percentage points in percentage of students who receive any SDO (Column I Panel A). ISS' drive this result; assigning more observations is associated with a statistically significant reduction of 0.31 percentage points in the percentage of students who receive ISS (Column I Panel B). The association with OSS is also statistically significant with a reduction of 0.18 percentage points (Column I Panel C). Finally, the association with EX is precise, near-zero, and statistically insignificant (Column I Panel D).

Results from Equation 2 are reported in column II and show that the findings are largely insensitive to unobserved time-invariant, between-school differences (i.e., school fixed effects). The association with any SDO increased slightly in magnitude to a 0.33 percentage point reduction (Column II Panel A). ISS results are nearly identical with a 0.30 percentage point decrease (column II Panel B). The association with OSS increases in magnitude to a 0.25 percentage points reduction (Column II Panel C), and the association with EX remains precise, near-zero, and statistically insignificant (Column II Panel D). The insensitivity of the results in column I to school fixed effects suggests that between-school endogeneity is not concerning.

Mediators. Results indicate that *Class Management* mediates the relationship between policy-assigned observations and ISS, but not the relationship with OSS. However, no evidence suggests that *Administrator Presence* or *Administrator Time* mediate any relationship (see Online Appendix C). Table 3 Column I reprints the unmediated relationships between observations assigned and ISS (Panel A) and OSS (Panel B), the two SDO-specific outcomes the evidence suggests are affected by policy-assigned observations; these results mean that there are unmediated relationships to explain. However, the *Class Management* mechanism can only explain the unmediated relationships if policy-assigned observations also affect *Class Management*. Indeed, Table 3 Column II shows that assigning schools 25 more observations is associated with an increase in the average environmental rating score of 0.03 units (0.08 SD). The final step in our investigation of the *Class Management* mediator concerns the extent to which the unmediated relationships in Column I change after we apply Equation 4. When mediated by *Management*, an increase in 25 policy-assigned observations is associated with a 0.25 percentage point decrease in ISS (Panel A Column III); the magnitude of the *Class-Management*-mediated relationship is 17% smaller than the main relationship (Panel A Column IV). However, the relationship between policy-assigned observations and OSS is effectively unchanged after adding the *Class Management* variable. These results suggest that *Class Management* mediates the relationship with ISS but not OSS.

Moderators. Moderation analyses yield two broad findings. First, none of the potential moderators affect the relationship with ISS, one moderates the OSS relationship, and every variable examined moderates the EX relationship. Second, relationships with OSS and EX strongly depend on the teacher experience moderator. None of the variables examined moderate relationships with ISS (see Online Appendix D) – we do not discuss these findings further.

The relationship with OSS does not depend on prior-year SDOs, while the relationship with EX does (Table 4 Panels A and B Column I). The evidence suggests that schools with lower proportions of expelled students from the prior year benefit more from policy-assigned observations than schools with higher proportions of expelled students (Table 4 Panel B Column I); however, these benefits are practically insignificant (negative coefficients represent benefits). Average teacher prior-year effectiveness scores moderate the relationship with EX but not OSS

(Table 4 Column II), like findings from the prior-year SDO moderation analysis. The data suggest that more policy-assigned observations reduce EX more if a school's average teacher has higher prior-year effectiveness scores than in schools where the average teacher's prior-year effectiveness score is lower (Table 4 Panel B Column II). Finally, more policy-assigned observations are associated with larger decreases in OSS and EX in schools where the average teacher has less experience than in schools where teachers have more years of experience (Table 4 Column III).

Sensitivity Tests. Local regression estimates are consistent in direction with the OLS and school fixed effect results; however, all local regression coefficients are statistically insignificant ($p > 0.05$). As observation assignments to the teachers whose prior-year effectiveness scores were just to either side of the 200 or 425 prior-year LOE thresholds increase, ISS, OSS, or EX decrease (Online Appendix Table B1 Panel A); however, the magnitude of the decline is less than half of the original estimate. Similar patterns exist with the ISS–only, OSS–only, and EX–only outcomes (Table B1 Panels B - D): the directions of the relationships are consistent with the OLS and school fixed effect results, the local regression estimates' magnitudes are at most half that of the OLS and school fixed effects results, and local regression standard errors are inflated, sometimes 100% larger than those in Table 2.

Formal sensitivity tests reveal that our inferences are insensitive to plausible OVs and that while implausible OVs can undo our inferences, they do not yield nonnegative coefficients, bolstering our confidence in the negative relationship between policy-assigned observations and SDOs. First, it is noteworthy that the school fixed effect model explains outcome and treatment variation remarkably well: 91% of the variation in the ISS/ OSS/ EX outcome and 92% of the variation in the number of policy-assigned observations. Thus, the number of plausibly threatening OVs is limited to those that could explain outcome and treatment variation independent of the variation explained by school and year fixed effects, the lagged outcome, two treatment determinants, and the rest of our rich covariates. Results in Table 5 further limit the set of plausibly threatening OVs. The prior-year SDO-benchmarked OV explains less than one-hundredth of the treatment residual variation and about 15% of residual outcome variation; however, an OV resembling this powerful predictor of residual outcome variation has virtually no effect on our coefficients or inferences. Although OVs explaining twice and thrice the variation of the prior-year SDO-benchmarked OV are assumedly implausible, such OVs still leave our inferences intact (Rows II, III). Similarly, our inferences are insensitive to the once, twice, and thrice-scaled teacher experienced-benchmarked OV, though years of experience was one of two variables that determined the number of policy-assigned observations (Rows IV-VI). Rows VII-IX reveal that while our inferences are insensitive to an OV benchmarked against the second determinant of policy-assigned observations, they are not immune to assumedly implausible OVs with twice and thrice its explanatory power. Although the estimates in Rows VIII and IX lose statistical significance, it is notable that these implausible OVs do not result in near-zero or positive coefficients.

Rothstein Falsification Tests. We attempt to falsify the findings from Equation 2 using "Rothstein falsification tests" (2010). Future policy-assigned observations cannot affect past SDOs causally. If Rothstein falsification tests detect relationships between future policy-assigned observations and past SDOs, the estimates generated by Equation 2 (school fixed effect model) are biased due to the endogenous sorting of students, teachers, or administrators into schools. We apply the falsification test by replacing $assigned_{st}$with $assigned_{s(t+1)}$; otherwise, the falsification test model is the same as Equation 2. Results from the Rothstein falsification tests suggest that endogenous sorting is not a concern as all estimates in Online Appendix Table B2 are precisely estimated, near-zero, and statistically insignificant.

Observations Received

Whereas an increase in the number of policy-assigned observations is associated with decreases in SDOs, an increase in the number of observations received is associated with increases in SDOs (Table 6 Columns I - II). For every 25 observations received, the percentage of students who receive any SDO is predicted to rise by a statistically significant 0.24 percentage points (Table 6 Panel A Column I). Again, this main result is driven by ISS'; an increase in observations received is associated with a significant 0.31 percentage point increase in ISS (Table 6 Panel B Column I). The association with OSS is small and not statistically significant (Table 6 Panel C Column I). Finally, the association with EX is precise and near-zero (Table 6 Panel D Column I). Results are insensitive to the inclusion of school fixed effects (Table 6 Column II), again suggesting that between-school endogeneity is not concerning.

2SLS results suggest that the schools assigned more observations also conduct more observations and that the policy-induced variation in observations received is associated with SDO reductions, consistent with our main findings in Table 2. First-stage results suggest that assigning schools 25 more observations is associated with an increase of approximately six observations conducted (0.24 * 25; Table 6 Column III *First Stage*); furthermore, all first-stage *F*-statistics confirm the strength of the instrument.[4] The second stage effectively estimates the association between SDOs and policy-induced observations *received* (i.e., the observations received as a result of policy rather than observations received due to administrator discretion). As 2SLS removes discretionary observation variation from the variation in observations

[4] *F*-statistics vary across panels in Column III because each regression uses a different prior-year 'outcome.'

received, second-stage results are effectively a confirmatory analysis of our main models from Table 2. Second-stage results suggest that policy-induced observations received are associated with decreases in SDOs or, in the case of EX, precisely estimated near-zero relationships (Table 6 Column III). Furthermore, an increase of 25 policy-induced observations received is associated with a substantial decrease in any SDO (Table 6 Panel A Column III), and the relationship with ISS drives this result again (Table 6 Panels B - D Column III). The magnitudes of the second-stage associations in Table 6 Column III exceed those in Table 2 because our primary models effectively estimate intent-to-treat associations, while 2SLS estimates treatment-on-the-treated estimates. Finally, the 2SLS results in Table 6 Column III are again insensitive to school fixed effects.[5]

Discretionary Observations. The results in Tables 2 and 6 raise questions about why associations with policy-assigned observations are negative while associations with observations received are positive; we respond by exploring the determinants of policy-assigned observations and observations received. Per Tennessee policy, $assigned_{st}$ is a function of teacher experience and prior-year effectiveness. We define $received_{st}$, the number of total observations received, as a function of $assigned_{st}$ and discretionary observations $discretionary_{st}$; where, $received_{st} = assigned_{st} + discretionary_{st}$. These definitions show that $received_{st}$ and $assigned_{st}$ differ due to $discretionary_{st}$.

We begin our exploration of determinants by graphically examining the relationship between $assigned_{st}$ and $discretionary_{st}$, which we operationalize as $received_{st} - assigned_{st}$. Figure 1 plots $discretionary_{st}$ against $assigned_{st}$ where circles represent cases

[5] Column IV applies the two-stage least-squares within estimator using Stata's *xtivreg* command, which does not report a first-stage F statistic.

when school-by-years conducted more observations than assigned by policy or the exact number assigned ($discretionary_{st} \geq 0$), and X's represent cases when school-by-years issued fewer observations than assigned by policy ($discretionary_{st} < 0$). We separate $discretionary_{st}$ into negative and non-negative cases if these two types of discretionary observations arise for different reasons. Graphically, there does not appear to be an unconditional relationship between policy-assigned observations and discretionary observations, nor does there appear to be a relationship between policy-assigned observations and discretionary observations conditional on whether $discretionary_{st}$ is negative or non-negative. More formally, the unconditional correlation between policy-assigned and discretionary observations is -0.37; the correlation between policy-assigned and discretionary observations conditional on ($discretionary_{st} \geq 0$) is 0.24; and the correlation between policy-assigned and discretionary observations conditional on ($discretionary_{st} < 0$) is 0.19. While these correlations suggest heterogeneity in the relationship between policy-assigned and discretionary observations, all the correlations are small, suggesting that the determinants of policy-assigned observations (i.e., teacher experience and prior-year effectiveness scores) are not related to the determinants of discretionary observations. We confirm this implication by regressing $discretionary_{st}$ on $experience_{st}$ and $LOE_{s(t-1)}$, the average teacher's years of experience and prior-year effectiveness score; the adjusted R^2 for this model is only 0.07. We suspected that prior-year SDOs might inform why administrators conduct discretionary observations and tested our suspicion by adding $SDO_{s(t-1)}$ as the third right-hand side variable for the $discretionary_{st}$ regression; the adjusted-R^2 for this model is still small at 0.13. We then add all the right-hand side observables from previous equations (X_{st}) and year fixed effects (α_t), which yields an adjusted-R^2 of 0.28. Finally, we add school fixed effects, which boosts the adjusted-R^2 to 0.49, but still explains less than half of the variation in discretionary observations. Ultimately, the data suggest that substantively different factors determine discretionary and policy-assigned observations.

Conclusions

We applied several regression analyses to multiple years of Tennessee administrative data to better understand the main, mediated, and moderated relationships between SDOs and the number of observations assigned to and conducted by schools. Our investigations yielded four key findings.

Policy-Assigned Observations May Reduce Widespread SDOs

The data repeatedly suggested that the percentage of students receiving at least one SDO decreased in schools that were assigned more teacher observations based on teacher years of experience and prior-year effectiveness scores. OLS compared schools to themselves and other observably similar schools, school fixed effects effectively compared schools to themselves over time, and local regressions only compared the teachers who fell just to either side of observation-assignment thresholds. Across the different comparisons made by each model, each found negative relationships between policy-assigned observations and SDOs, though local regression estimates were estimated imprecisely. Furthermore, Rothstein falsification tests could not falsify school fixed effect results, and those results were insensitive to omitted variables benchmarked against the most potent determinants of SDOs and policy-assigned observations. Although implausible OVs with twice the explanatory power of one policy determinant pushed estimates to statistical insignificance, undoing our inferences, those OVs did not result in nonnegative coefficients. We conclude that more policy-assigned observations plausibly reduce SDOs.

Notably, the statistically significant relationships between SDOs and observations are meaningful. Regressions estimated changes in SDOs that were associated with assigning schools

25 more teacher observations. If the average school - which is assigned just over 63 observations – was assigned 25 more observations, this would be a 40% increase, resulting in roughly 88 total assigned observations. At face value, 25 more observations may seem too costly in terms of administrator time, psychological burdens, and observation process quality. However, recent Tennessee-wide panel data analyses find that administrators who conduct additional observations do not report being more stressed or burdened (Hunter & Rodriguez, 2021); yet, Hunter and Rodriguez also found that administrators coped with observation demands by abbreviating observation length and pre- and post-observation conferences. Despite the potential losses in observation process quality related to increasing observations, our data suggest that the net effects may result in practically substantial changes in SDOs. OLS and school fixed effect models estimated an approximately 0.30-point reduction in the percentage of students receiving any SDO. As the average Tennessee school enrolls about 430 students, a 0.30-percentage-point reduction amounts to about one less student receiving an SDO in the average school each year. While this may seem insignificant, prior work by Carrell and colleagues (2018) suggests otherwise; foregoing exposure to a disruptive classroom peer can increase earnings at ages 24 - 28 by 3 percent. Back-of-the-envelope calculations suggest that if one student in each of Tennessee's roughly 2000 schools did not receive an SDO, the total increase in long-term income would be over $2 million.

The data also suggest that observation assignments may not affect all SDOs equally, as we hypothesized. Although our hypotheses and prior work implied that assigned observations might affect SDOs (Liebowitz et al., 2022; Rodriguez & Welsh, 2022), the observations we examined were not purposefully designed for SDO reduction; as such, we characterized observation assignments as a light-touch SDO intervention. Furthermore, we hypothesized that it is easier for a light-touch intervention to change the factors causing ISS than the factors driving OSS, which we hypothesized is easier than changing the factors that determine EX, consistent with previous research (Lacoe & Steinberg, 2019). Indeed, the approximately 0.30-point SDO reduction was driven by an approximate 0.30-point decrease in the percentage of students receiving at least one ISS and an approximate 0.20-point reduction in the percentage of students receiving at least one OSS, and the percentage of students receiving at least one EX was insensitive to the number of observations assigned to schools by policy, on average. As the percentage of students receiving ISS exceeds the percentage receiving OSS, which exceeds the percentage receiving EX, ISS reductions affect more students than OSS reductions and far more than EX reductions. Ultimately, if assigning observations reduces SDOs, the effects are practically meaningful and affect the most widespread SDOs.

Policy-Assigned Observations May Reduce ISS by Improving Class Management

The data suggested that only one of the three mechanisms investigated mediates the relationship between observation assignment and ISS or OSS, and that mechanism affects ISS only.[6] We hypothesized that observation assignments might affect SDOs by improving *Class Management*, increasing mollifying *Administrator Presence* in classrooms, and reducing *Administrator Time* spent on student discipline issues. None of the evidence supported our hypotheses regarding *Administrator Time.*

Observation assignments were related to the *Administrator Presence* variable but not in the direction expected, and the data suggested that the variable was not a mediator. TES teacher respondents in the same school over time reported that observation-induced administrator presence was more likely to disrupt instruction during the years when the school was assigned

[6] As OLS and school fixed effect models generated precisely estimated near-zero associations with EX, there were not EX relationships to mediate.

more observations. On the contrary, we hypothesized that teachers would welcome more frequent administrator presence in classrooms to manage student behavior better. Although assigning more observations is associated with more disruptive administrator presence, these disruptions did not explain any relationship between policy-assigned observations and SDOs, suggesting that the *Administrator Presence* variable is not a mediator.

However, more observation assignments were associated with *Class Management* improvements, which mediated almost 20% of the relationship between observation assignments and ISS. Assigning schools more observations may provide evaluators with multiple opportunities to discuss and recommend improvements regarding classroom management; indeed, when a school was assigned 25 more observations, the average teacher's classroom management improved by 0.03 points (0.08 SD). Although an 0.03-unit improvement is not practically significant (it moves a school in the *Class Management* distribution from the 50th to the 53rd percentile) and mediated virtually none of the OSS relationship, it explained nearly 20% of the ISS relationship.

The *Class Management* mediation analyses underscore the malleability of ISS and suggest that a sizeable percentage of ISS' may arise from classroom-based student misbehavior or how teachers handle ISS-related misbehaviors. First, the evidence suggests that a light-touch intervention (observation assignment) may improve classroom management slightly, but slight improvements may be enough to reduce ISS appreciably. As ISS' are based on mild (perceived) student misbehaviors (Rodriguez & Welsh, 2022), reducing ISS may only require minor changes to school disciplinary processes. Alternatively, reducing OSS and EX' may require more intensive interventions. Second, the RW framework (2022) argued that educators respond to perceived student misbehavior via classroom or office management. Office management always results in office referrals and may result in SDOs. However, classroom-managed responses by teachers do not lead to SDOs; instead, they lead to non-exclusionary outcomes (e.g., parent phone calls). If observation assignments improve classroom management, this may reduce the occurrence of ISS-based misbehaviors (teachers manage the classroom better) or increase the odds that teachers respond to ISS-based misbehaviors without resorting to office management, diverting the school disciplinary process away from exclusionary outcomes like ISS'.

OSS and EX Relationships Depend on School Characteristics Examined, ISS Relationships Do Not

We examined three policy- and conceptually-relevant moderators: prior-year SDOs, average teacher years of experience, and average teacher prior-year effectiveness scores. None of these three variables moderated relationships with ISS. On the one hand, the lack of heterogeneity in ISS relationships may mean that observation assignments work equally well across the school characteristics examined. Alternatively, the results also mean that observation assignments do not affect ISS equitably; otherwise, we would have observed larger benefits in schools with high prior-year SDOs, which would shrink the gap between schools with higher percentages of students receiving in-school suspensions and those with lower percentages.

Only average teacher years of experience moderated the OSS relationship. Although we are unaware of research documenting the returns to teaching experience regarding SDOs, we assume that teachers become more effective classroom managers with time. Thus, early-career teachers need to improve more than later-career teachers; this logic also underpins the assignment of more observations to early-career Tennessee teachers. Moderation analyses suggest that the OSS reductions associated with more observation assignments become smaller in magnitude as the school-level average teacher gains experience. For example, the data suggest that schools at the 25th percentile of the average teacher years of experience distribution (10.4 years) may see an 0.35-point OSS reduction. In comparison, schools at the 75th percentile (14 years) are predicted to see an 0.21-point OSS reduction. Moreover, schools in the bottom 10% of the distribution of average teacher experience (5 - 8.67 years) may see OSS reductions between 0.57 and 0.42. Thus, schools with the least experienced teachers benefit most from being assigned more observations by state policy.

Although observation assignments are unrelated to EX, on average, the relationship is correlated with the three moderators examined. Negative associations between observation assignments and EX are largest in schools with lower percentages of students receiving EX in the prior year, where the average teacher has fewer years of experience, and where the average teacher has higher prior-year effectiveness scores. Regarding the prior-year EX moderator, the data suggest that the only schools in which more observation assignments may reduce EX are schools where less than one percent of students received at least one EX in the prior year. However, the bottom 85% of schools had zero EX in the prior year, suggesting that assigning schools 25 more observations may be able to reduce EX in a substantial number of settings.

The data suggest that the EX reductions associated with more observation assignments shrink as the average teacher gains experience, meaning that teacher experience moderates associations with the two SDO categories based on the most severe student misbehaviors. However, the experience moderator shrinks the negative EX associations toward zero at a significantly faster rate than it shrinks the negative OSS relationships toward zero. For example, EX relationships are only negative for schools in the bottom five percent of the experience distribution, but OSS relationships are negative for all schools but the top five percent. The limited scope of negative EX relationships limits the practical significance of the average teacher experience moderator, although it is statistically significant.

Finally, the data suggest that more observation assignments may reduce EX in schools where the average teacher had higher prior-year effectiveness scores. Indeed, EX associations are negative in schools where the average teacher's prior-year effectiveness score exceeded 350, as did schools in the upper 88% of the effectiveness moderator distribution.

Briefly, we conclude that assigning schools more observations associates with ISS similarly across the school settings examined, and a school's average teacher's years of experience moderates the OSS relationship substantially. The data also suggest several school settings where more policy-assigned observations may reduce EX as the three variables examined moderated the EX relationship; however, prior-year EX and average teacher prior-year effectiveness scores are the only practically significant moderators. Finally, while policy-assigned observations may reduce EX in most schools, where less than one percent of students were expelled in the prior year or the average teacher received a relatively high prior-year effectiveness score, the associations represent Matthew effects. Schools with higher concentrations of effective teachers and lower concentrations of severe misbehavior benefit from policy-assigned observations more than less-advantaged schools.

Discretionary and Policy-Assigned Observations Arise for Different Reasons and Affect SDOs Differently

Many schools deviate from the number of observations assigned, which we characterized as discretionary observations, often by conducting more observations than assigned; the average school conducts approximately 22 more observations than assigned. Although two teacher human capital measures determine the number of observations assigned, neither teacher nor administrator human capital measures nor any observable student, teacher, or administrator characteristic nor unobserved between-school differences explain most of the variation in discretionary observations.

While several factors might account for a sizeable fraction of the unexplained variation in discretionary observations, we argue that reverse causation is the most plausible. When teachers handle (perceived) student misbehavior via RW's (2022) "class management" mechanism, these cases do not appear in the administrative data. However, schools may respond to such cases by conducting more observations than assigned to mitigate (perceived) misbehaviors, whether they result in non-exclusionary outcomes or SDOs. If unobserved class-managed discipline occurrences also correlate positively with the office-managed occurrences that result in SDOs, this will introduce positive bias into the estimated relationship between SDOs and observations received. We also characterize situations when schools anticipate higher SDOs and respond by conducting more observations as reverse causation. Our arguments imply that the SDOs in year *t* may explain discretionary observations in year *t.* In unreported regressions, we added the SDOs in year *t* as a right-hand side variable in the model explaining discretionary observation variation based on all observables, year fixed effects, and school fixed effects. A one percentage point increase in SDOs in year *t* is statistically significantly associated with an increase of 0.25 discretionary observations in total; however, the increase is practically insignificant, and the new model still explains less than half of the variation in discretionary observations (adjusted-R^2 0.49). Notwithstanding the plausible endogeneity in estimated relationships between SDOs and observations received, SDOs are higher in schools conducting more observations, no matter the source of endogeneity. In conclusion, the data suggest that basing observations on teacher experience and prior-year effectiveness may reduce SDOs, while observations arising for unobserved reasons may increase them.

Limitations

This study may suffer from three broad limitations. First, the estimates may not capture precise causal estimates. Although evidence from mediation analyses, local regressions, and Rothstein falsification tests are consistent with causal interpretations, we are unsure if our estimates are strictly causal. Instead, we argue that associations with the number of observations assigned might approximate causal estimates. We suspect that associations with observations received are biased. Second, the RW framework (2022) suggests more mechanisms that might explain the relationships, or lack thereof, than the mechanisms we explored. Future work might apply different operationalizations of the mechanisms we explored and consider additional mechanisms to provide further insights into teacher evaluation and school disciplinary processes. Finally, we do not assume that our findings generalize to other settings. Relatedly, the conclusions based on policy-assigned observations may not generalize to observations not assigned because of teacher experience and prior-year performance.

Implications

The evidence implies that policymakers might increase the number of observations assigned to schools, but only if teacher experience and prior-year performance determine the number assigned. While our analyses suggest that assigning observations based on these two determinants may lower SDOs, discretionary observations that arise for unknown reasons were not associated with SDO reductions. In Tennessee, where observations are already assigned based on experience and prior-year performance, policymakers might increase the number of observations assigned to certain teacher groups. For example, policymakers might assign teachers with the lowest prior-year performance five observations instead of four or moderately effective teachers with more than four years of experience three observations instead of two.

Despite the seemingly minor changes in SDOs that 25 more experience-and-effectiveness-based observations may induce, prior work suggests that the reductions are meaningful and practically significant. Although ISS reductions drive the SDO reductions, moderation analyses suggest that experience-and-effectiveness-based observation assignments may lead to substantial OSS and EX reductions in most schools. Thus, experience-and-effectiveness-based observations may represent a tool that policymakers can use to reduce mild to severe (perceived) student misbehaviors resulting in exclusionary outcomes in most school settings. Additionally, assigning schools 25 more observations is feasible; indeed, the typical school already conducts 22 observations above the number assigned.

Moderation analyses also suggest that policymakers might assign relatively more experience-and-effectiveness-based observations to schools with higher concentrations of less effective teachers or higher concentrations of expelled students, characteristics observable to state and local policymakers, if they wish to reduce EX using observations. Alternatively, or in tandem, policymakers might assign additional support (e.g., executive or instructional coaches) to these schools. Mediation analyses suggest that additional support might focus on catalyzing the positive effects of assigned observations on classroom management; even small increases in classroom management skills may yield substantial SDO reductions.

Discretionary observation analyses suggest that administrators may need additional support regarding why they conduct additional observations. Schools deviate from the number of observations assigned for unknown and potentially counterproductive reasons. Professional development workshops, regular communication from education agencies, or light-touch executive coaching might help administrators better use their discretion to conduct SDO-reducing observations.

Finally, there are several implications regarding theory and research. First, the negative associations with experience-and-effectiveness-based observations and positive associations with observations that arise for different reasons support a core tenet of the strategic management of human capital theory (Odden & Kelly, 2008) and the theory of action framing teacher evaluation (Archer et al., 2016; Donaldson, 2021) - educational leaders and policymakers can meaningfully improve important student outcomes when adopting evaluation-related interventions based on teacher human capital measures. A growing body of work suggests that the introduction of teacher evaluation reforms did not improve student outcomes, on average (Bleiberg et al., 2021; Hunter & Bowser, 2021; Liebowitz et al., 2022; Steinberg & Sartain, 2015; Taylor & Tyler, 2012). In addition to the findings herein, prior work also finds that educational leaders deviate from human-capital-informed decision-making (Donaldson & Woulfin, 2018; Hunter, 2020; Hunter & Ege, 2021; Marsh et al., 2017; Rodriguez & Hunter, 2021). We suspect that reformed teacher evaluation systems might improve multiple student outcomes significantly if educational leaders reliably implemented these systems based on student performance, and teacher and administrator human capital. Policy or professional learning opportunities (e.g., coaching) may tighten the link between human capital measures and evaluation interventions for student benefit.

References

American Institutes for Research. (2016). *Databases on State Teacher and Principal Evaluation Policies (STEP Database and SPEP Database)*. http://resource.tqsource.org/stateevaldb/Compare50States.aspx

Archer, J., Cantrell, S., Holtzman, S., Joe, J., Tocci, C., & Wood, J. (2016). *Better feedback for better teaching: A practical guide to improving classroom observations* (1st ed.). Jossey-Bass.

Bleiberg, J., Brunner, E., Harbatkin, E., Kraft, M. A., & Springer, M. G. (2021). *The Effect of Teacher Evaluation on Achievement and Attainment: Evidence from Statewide Reforms* (Working Paper 21–496; EdWorkingPaper). Annenberg Institute at Brown University. https://www.edworkingpapers.com/ai21-496

Carrell, S., Hoekstra, M., & Kuka, E. (2018). The Long-Run Effects of Disruptive Peers. *American Economic Review*, *108*(11). https://doi.org/10.3386/w22042

Chetty, R., Friedman, J. N., & Rockoff, J. E. (2014). The Long Term Impacts of Teachers: Teacher Value Added and Student Outcomes in Adulthood. *American Economic Review*, *104*(9), 2633–2679.

Cinelli, C., & Hazlett, C. (2020). Making Sense of Sensitivity: Extending Omitted Variable Bias. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, *82*(1), 39–67. https://doi.org/10.1111/rssb.12348

de Barros, A. (2019). Evaluating Teacher Evaluation: Evidence from Chile. *Organization of Schools and Systems & Education in Global Contexts*. Society for Research in Educational Effectiveness, Washington, DC.

Demerouti, E., Bakker, A. B., Nachreiner, F., & Schaufeli, W. B. (2001). The job demands-resources model of burnout. *Journal of Applied Psychology*, *86*(3), 499–512. https://doi.org/10.1037/0021-9010.86.3.499

Donaldson, M. L. (2021). *Multidisciplinary Perspectives on Teacher Evaluation: Understanding the Research and Theory* (1st ed.). Routledge.

Donaldson, M. L., & Papay, J. P. (2014). Teacher Evaluation for Accountability and Development. In *Handbook of Research in Education Finance and Policy* (pp. 190–209). Routledge.

Donaldson, M. L., & Woulfin, S. (2018). From Tinkering to Going "Rogue": How Principals Use Agency When Enacting New Teacher Evaluation Systems. *Educational Evaluation and Policy Analysis*, *40*(4), 531–556. https://doi.org/10.3102/0162373718784205

Grissom, J. A., Blissett, R. S. L., & Mitani, H. (2018). Evaluating School Principals: Supervisor Ratings of Principal Practice and Principal Job Performance. *Educational Evaluation and Policy Analysis*, *40*(3), 446–472. https://doi.org/10.3102/0162373718783883

Huang, T., Hochbein, C., & Simons, J. (2020). The relationship among school contexts, principal time use, school climate, and student achievement. *Educational Management Administration & Leadership*, *48*(2), 305–323. https://doi.org/10.1177/1741143218802595

Hunter, S. B. (2019). *The Effects of More Frequent Observations on Student Achievement Scores* (2019–04; Strengthening Tennessee's Education Labor Market). Tennessee Education Research Alliance. https://peabody.vanderbilt.edu/TERA/files/TERA_Working_Paper_2019-04.pdf

Hunter, S. B. (2020). The Unintended Effects of Policy-Assigned Teacher Observations: Examining the Validity of Observation Scores. *AERA Open*, *6*(2). https://doi.org/10.1177/2332858420929276

Hunter, S. B. (2022). High-leverage teacher evaluation practices for instructional improvement. *Educational Management Administration & Leadership*, 174114322211129. https://doi.org/10.1177/17411432221112995

Hunter, S. B. (2023). Do You Mean What I Mean? Comparing Teacher Performance Self-Scores and Evaluator-Generated Scores. *Journal of Education Human Resources*, *41*(2), 210–250. https://doi.org/10.3138/jehr-2020-0026

Hunter, S. B., & Bowser, K. (2021). Identifying the Effects of Next-Generation Teacher Evaluation on Student Achievement in Rural Districts: Evidence from Missouri. *11.03 Teacher Evaluation Systems. Educator Preparation, Professional Development, Performance, and Evaluation*. Association for Education Finance and Policy Annual Conference, Virtual. https://education.gmu.edu/assets/docs/educational_leadership/HunterBowser_Introducing.pdf

Hunter, S. B., & Ege, A. (2021). Linking Student Outcomes to School Administrator Discretion in the Implementation of Teacher Observations. *Educational Administration Quarterly*, *57*(4), 607–640. https://doi.org/10.1177/0013161X211003134

Hunter, S. B., & Kho, A. (2023). The Effects of Teacher Evaluation Policy on Student Achievement and Teacher Turnover: Leveraging Teacher Accountability and Teacher Development. *Journal of Education Human Resources*. Advance online publication. https://doi.org/10.3138/jehr-2023-0040

Hunter, S. B., & Rodriguez, L. A. (2021). Examining the demands of teacher evaluation: Time use, strain and turnover among Tennessee school administrators. *Journal of Educational Administration*, *59*(6), 739–758. https://doi.org/10.1108/JEA-07-2020-0165

Hunter, S. B., & Springer, M. G. (2022). Performance Feedback, Human Capital, and Teacher Performance: A Mixed-Methods Analysis. *Educational Evaluation and Policy Analysis*, *44*(3), 380–403. https://doi.org/10.3102/01623737211062913

Jackson, C. K. (2018). What Do Test Scores Miss? The Importance of Teacher Effects on Non–Test Score Outcomes. *Journal of Political Economy*, *126*(5), 2072–2107.

Johnson, S. M., & Fiarman, S. E. (2012). The Potential of Peer Review. *Educational Leadership*, 7.

Kraft, M. A., & Gilmour, A. F. (2016). Can Principals Promote Teacher Development as Evaluators? A Case Study of Principals' Views and Experiences. *Educational Administration Quarterly*, *52*(5), 711–753. https://doi.org/10.1177/0013161X16653445

Kraft, M. A., Papay, J. P., & Chi, O. (2020). Teacher Skill Development: Evidenced from Performance Ratings by Principals. *Journal of Policy Analysis and Management*, *39*(2), 315–347.

Lacoe, J., & Steinberg, M. P. (2019). Do Suspensions Affect Student Outcomes? *Educational Evaluation and Policy Analysis*, *41*(1), 34–62. https://doi.org/10.3102/0162373718794897

Liebowitz, D. D., Porter, L., & Bragg, D. (2022). The Effects of Higher-Stakes Teacher Evaluation on Office Disciplinary Referrals. *Journal of Research on Educational Effectiveness*, *15*(3), 475–509. https://doi.org/10.1080/19345747.2021.2015496

Liu, J., & Loeb, S. (2021). Engaging Teachers: Measuring the Impact of Teachers on Student Attendance in Secondary School. *Journal of Human Resources*, *56*(2), 343–379. https://doi.org/10.3368/jhr.56.2.1216-8430R3

Marsh, J. A., Bush-Mecenas, S., Strunk, K. O., Lincove, J. A., & Huguet, A. (2017). Evaluating Teachers in the Big Easy: How Organizational Context Shapes Policy Responses in New Orleans. *Educational Evaluation and Policy Analysis*, *39*(4), 539–570. https://doi.org/10.3102/0162373717698221

National Council on Teacher Quality. (2019). *NCTQ: Yearbook: Frequency of Evaluation and Observation*. National Council on Teacher Quality (NCTQ). https://www.nctq.org/yearbook/national/Frequency-of-Evaluation-and-Observation-95#undefined

Neumerski, C. M. (2013). Rethinking Instructional Leadership, a Review: What Do We Know About Principal, Teacher, and Coach Instructional Leadership, and Where Should We Go From Here? *Educational Administration Quarterly*, *49*(2), 310–347. https://doi.org/10.1177/0013161X12456700

Odden, A., & Kelly, J. A. (2008). *Strategic Management Of Human Capital In Public Education*. Strategic Management of Human Capital.

Osborne, Jr., A. G. (2001). Discipline of special-education students under the Individuals with Disabilities Education Act. *Fordham Urban Law Journal*, *29*(513).

Papay, J. P. (2012). Refocusing the Debate: Assessing the Purposes and Tools of Teacher Evaluation. *Harvard Educational Review*, *82*(1), 123–141. https://doi.org/10.17763/haer.82.1.v40p0833345w6384

Papay, J. P., & Kraft, M. A. (2013). Productivity returns to experience in the teacher labor market: Methodological challenges and new evidence on long-term career improvement.

Journal of Public Economics, *130*, 105–119. https://doi.org/10.1016/j.jpubeco.2015.02.008

Phipps, A., & Wiseman, E. A. (2021). Enacting the Rubric: Teacher Improvements in Windows of High-Stakes Observation. *Education Finance and Policy*, *16*(2), 283–312. https://doi.org/10.1162/edfp_a_00295

Rigby, J. G. (2015). Principals' Sensemaking and Enactment of Teacher Evaluation. *Journal of Educational Administration*, *53*(3), 374–392. https://doi.org/10.1108/JEA-04-2014-0051

Rodriguez, L. A., & Hunter, S. B. (2021). Making Do: Why Do Administrators Retain Low-Performing Teachers? *Educational Researcher*, *50*(9), 673–676. https://doi.org/10.3102/0013189X211039450

Rodriguez, L. A., & Welsh, R. O. (2022). The Dimensions of School Discipline: Toward a Comprehensive Framework for Measuring Discipline Patterns and Outcomes in Schools. *AERA Open*, *8*, 233285842210836. https://doi.org/10.1177/23328584221083669

Rosenbaum, P. R., & Rubin, D. B. (1983). Assessing Sensitivity to an Unobserved Binary Covariate in an Observational Study with Binary Outcome. *Journal of the Royal Statistical Society Series B*, *45*(2), 212–218.

Rothstein, J. (2010). Teacher Quality in Educational Production: Tracking, Decay, and Student Achievement. *The Quarterly Journal of Economics*, *125*(1), 175–214. https://doi.org/10.1162/qjec.2010.125.1.175

Skiba, R. J., Chung, C.-G., Trachok, M., Baker, T. L., Sheya, A., & Hughes, R. L. (2014). Parsing Disciplinary Disproportionality: Contributions of Infraction, Student, and School Characteristics to Out-of-School Suspension and Expulsion. *American Educational Research Journal*, *51*(4), 640–670. https://doi.org/10.3102/0002831214541670

Stecher, B. M., Garet, M. S., Hamilton, L. S., Steiner, E. D., Robyn, A., Poirier, J., Holtzman, D., Fulbeck, E. S., Chambers, J., & Brodziak de los Reyes, I. (2016). *Improving Teaching Effectiveness*.

Steinberg, M. P., & Donaldson, M. L. (2016). The New Educational Accountability: Understanding the Landscape of Teacher Evaluation in the Post-NCLB Era. *Education Finance and Policy*, *11*(3). https://doi.org/10.1162/EDFP_a_00186

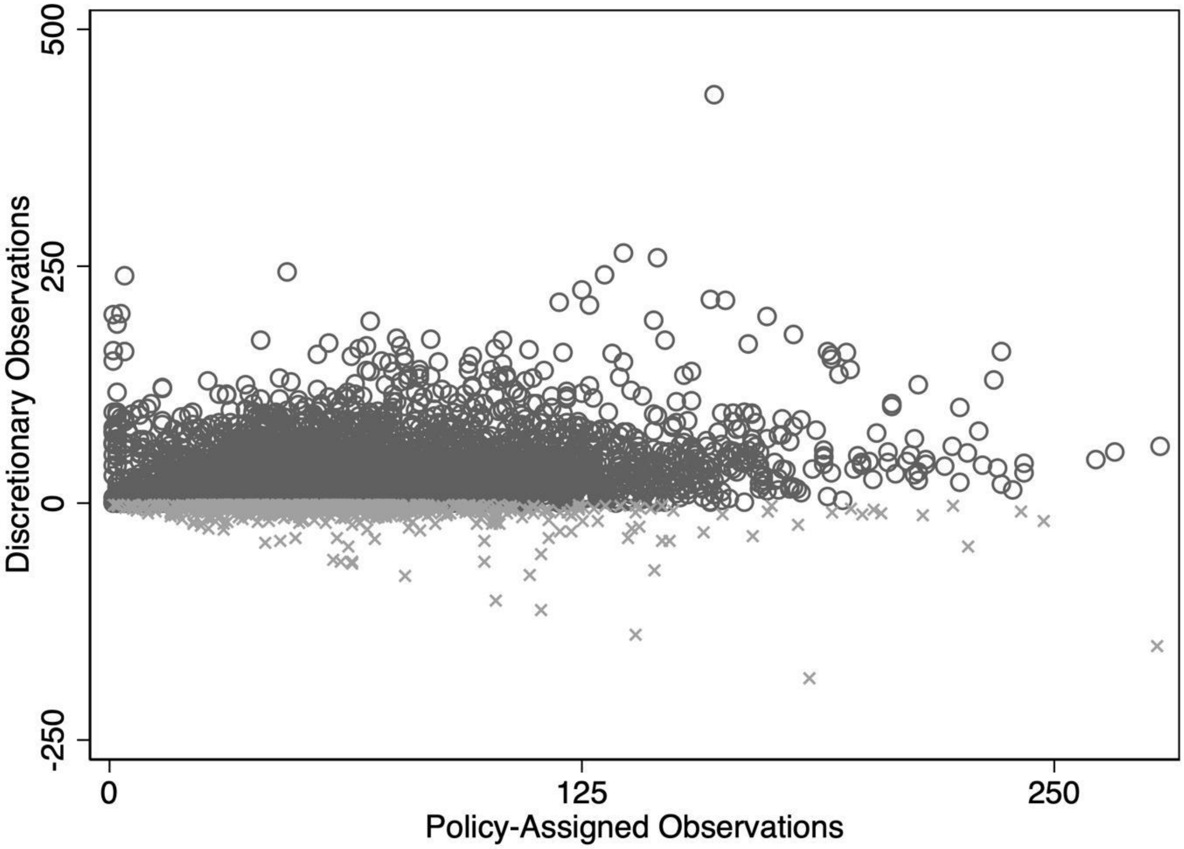
Steinberg, M. P., & Sartain, L. (2015). Does teacher evaluation improve school performance? Experimental evidence from Chicago's Excellence in Teaching Project. *Education Finance and Policy*, *10*(4), 535–572. https://doi.org/10.1162/EDFP_a_00173

Taylor, E. S., & Tyler, J. H. (2012). The Effect of Evaluation on Teacher Performance. *American Economic Review*, *102*(7), 3628–3651. https://doi.org/10.1257/aer.102.7.3628

Welsh, R. O., & Little, S. (2018). The School Discipline Dilemma: A Comprehensive Review of Disparities and Alternative Approaches. *Review of Educational Research*, *88*(5), 752–794. https://doi.org/10.3102/0034654318791582

Figure 1

Scatterplot: Discretionary Observations and Policy-Assigned Observations



Notes: School-by-years are the unit of analysis. Plots total discretionary observations (i.e., total observations received – total number of observations assigned by policy) against total policy-assigned observations (both variables expressed in 25s). Circles represent cases when schools issued more observations than assigned by policy; X's represent cases issued fewer observations than assigned by policy. The correlation between discretionary and policy-assigned observations is -0.37; the correlation between discretionary and policy-assigned observations in schools where teachers discretionary observations $\geq$ 0 is 0.24; and, the correlation between discretionary and policy-assigned observations in schools where teachers discretionary observations < 0 is 0.19.

Table 1

Sample Characteristics

	Mean	Standard Deviation
Percentage of Students Receiving Any Suspension/ Expulsion At Least Once	8.45	10.2
Percentage of Students Receiving In-School Suspension At Least Once	6.09	9.0
Percentage of Students Receiving Out-of-School Suspension At Least Once	3.77	5.5
Percentage of Students Expelled At Least Once	0.07	0.3
Observations Assigned	63.25	35.5
Observations Received	85.81	49.9
N(Schools)	1369	
N(School-by-Years)	6329	

Notes: School-by-years are the unit of analysis.

Table 2

Relationships Between Discipline Outcomes and Observations Assigned

	I	II
Panel A. DV = Percentage of Students Receiving Any Suspension/ Expulsion		
Obs Assigned (25s)	-0.30***	-0.33**
	(0.09)	(0.12)
Panel B. DV = Percentage of Students Receiving In-School Suspension		
Obs Assigned (25s)	-0.31**	-0.30*
	(0.10)	(0.13)
Panel C. DV = Percentage of Students Receiving Out-of-School Suspension		
Obs Assigned (25s)	-0.18**	-0.25***
	(0.06)	(0.07)
Panel D. DV = Percentage of Students Expelled		
Obs Assigned (25s)	0.01	0.00
	(0.01)	(0.01)
School FE		X
N(School-by-Years)	6329	6329

Notes: School-by-years are the unit of analysis. The outcomes are percentages ranging from zero to 100. The independent variable, the number of observations assigned, is scaled by 25s; thus, coefficients represent the change in an outcome associated with an increase of 25 policy-assigned observations. Each model controls for year fixed effects and the prior-year 'outcome;' average teacher and administrator prior-year composite effectiveness scores; the proportions of students, teachers, and administrators who are white, black, and female; the proportions of students who are ELL, SPED, and economically disadvantaged; average teacher and administrator years of experience and level of education; the number of teachers in each school; and the ratio of teachers to evaluators in each school. Standard errors are clustered at the school level. * *p* < 0.05, ** *p* < 0.01, *** *p* < 0.001.

Table 3

Classroom Management Mediation Analysis: Relationships Between Policy-Assigned Observations and Discipline Outcomes

	I	II	III	IV
	Main Relationship	Mediator Relationship	Mediated Main Relationship	% change in main relationship
Panel A. DV = In-School Suspensions				
Obs Assigned (25s)	-0.30*	0.03***	-0.25	+17%
	(-0.57, -0.04)	(0.02, 0.05)	(-0.51, 0.01)	
Panel B. DV = Out-of-School Suspensions				
Obs Assigned (25s)	-0.25***	0.03***	-0.26***	-2%
	(-0.39, -0.12)	(0.02, 0.05)	(-0.39, -0.13)	
N(School-by-Year)	6204	6204	6204	

Notes: School-by-years are the unit of analysis. Standard errors are clustered at the school level; 95% confidence intervals in parentheses. The discipline outcomes are percentages ranging from zero to 100; the environment rating variable ranges from 1.75 to 5, with a 4.23 mean and 0.38 standard deviation. The number of observations assigned are scaled by 25s; thus, coefficients represent association with an increase of 25 observations. Column I lists unmediated results using the same sample as the mediation analyses. Column II treats the mediator as the 'outcome.' Column III uses the same outcome and right hand-side variables as column I but adds the mediator as a right hand-side variable. Each model controls for year and school fixed effects and the prior-year 'outcome;' average teacher and administrator prior-year composite effectiveness scores; the proportions of students, teachers, and administrators who are white, black, and female; the proportions of students who are ELL, SPED, and economically disadvantaged; average teacher and administrator years of experience and level of education; the number of teachers in each school; and the ratio of teachers to evaluators in each school. * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$.

Table 4

Moderated Relationships Between Policy-Assigned Observations and Discipline Outcomes

	I	II	III
Panel A. DV = Percentage of Students Receiving Out-of-School Suspension			
Obs Assigned (25s)	-0.24*	-0.01	-0.77**
	(0.09)	(0.26)	(0.29)
Obs Assigned (25s) * Prior-Year OSS	-0.004		
	(0.004)		
Obs Assigned (25s) * Avg Tch Prior-Year LOE (50s)		-0.03	
		(0.04)	
Obs Assigned (25s) * Avg Tch Yrs Experience			0.04*
			(0.02)
Panel B. DV = Percentage of Students Expelled			
Obs Assigned (25s)	-0.01*	0.07**	-0.06*
	(0.01)	(0.03)	(0.02)
Obs Assigned (25s) * Prior-Year Expulsions	0.01***		
	(0.02)		
Obs Assigned (25s) * Avg Tch Prior-Year LOE (50s)		-0.01**	
		(0.004)	
Obs Assigned (25s) * Avg Tch Yrs Experience			0.01**
			(0.002)
N(School-by-Years)	6329	6329	6329

Notes: School-by-years are the unit of analysis. Each column-by-panel presents results from a separate regression. The outcomes are percentages ranging from zero to 100. The independent variable, the number of observations assigned, is scaled by 25s; thus, coefficients represent the change in an outcome associated with an increase of 25 policy-assigned observations. The average teacher years of experience and prior-year SDOs are scaled by ones; average teacher prior-year LOE is scaled by 50. Each model includes year and school fixed effects and controls for the prior-year 'outcome;' average teacher and administrator prior-year composite effectiveness scores; the proportions of students, teachers, and administrators who are white, black, and female; the proportions of students who are ELL, SPED, and economically disadvantaged; average teacher and administrator years of experience and level of education; the number of teachers in each school; and the ratio of teachers to evaluators in each school. Standard errors are clustered at the school level. * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$.

Table 5

Sensitivity of Inferences to Benchmarked Omitted Variables

		I	II	III
		$R^2_{Pol\ Obs\sim OV\mid X}$	$R^2_{SDO\sim OV\mid X,Pol\ Obs}$	Coef (SE)
I	1.00x Prior-Year SDOs	< 0.0001	0.1452	-0.31** (0.10)
II	2.00x Prior-Year SDOs	0.0001	0.2904	-0.29** (0.09)
III	3.00x Prior-Year SDOs	0.0001	0.4356	-0.28***](0.08)
IV	1.00x Avg Teacher Years of Exp	< 0.0001	0.0011	-0.33** (0.11)
V	2.00x Avg Teacher Years of Exp	< 0.0001	0.0022	-0.33** (0.11)
VI	3.00x Avg Teacher Years of Exp	< 0.0001	0.0033	-0.33** (0.11)
VII	1.00x Avg Teacher Prior-Year LOE	0.1308	0.0004	-0.27* (0.12)
VIII	2.00x Avg Teacher Prior-Year LOE	0.2616	0.0009	-0.19 (0.13)
IX	3.00x Avg Teacher Prior-Year LOE	0.3924	0.0015	-0.10 (0.14)
	N(School-by-Years)			6329

Notes: Models apply Equation 2 (school fixed effect model). *Pol Obs* represents treatment, *OV* the hypothetical omitted variables, and *X* all righthand-side variables from Equation 2 excluding *Pol Obs*. $R^2_{Pol\ Obs\sim OV\mid X}$ represents the proportion of explained residual variation in *Pol Obs*. $R^2_{SDO\sim OV\mid X,Pol\ Obs}$ represents the proportion of explained residual variation in the *SDO* outcome. "Coef" is the estimated treatment effect if Equation 2 controlled for *OV*. * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$.

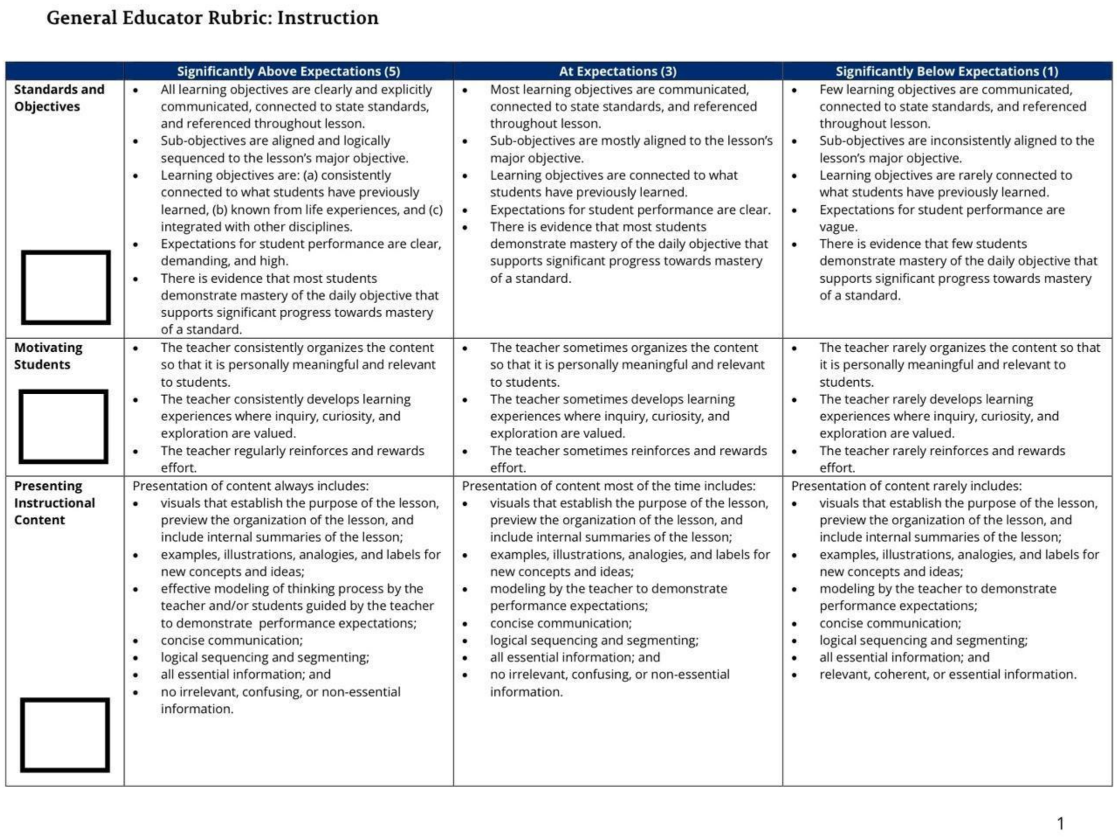
Table 6

Relationships Between Discipline Outcomes and Observations Received

	I	II	III	IV
Panel A. DV = Percentage of Students Receiving Any Suspension/ Expulsion				
Obs Received (25s)	0.24***	0.28*	-1.23**	-1.83*
	(0.07)	(0.11)	(0.41)	(0.79)
First Stage				
Obs Assigned (25s)			0.24***	0.18***
			(0.04)	(0.04)
F-Stat			42.66***	
Panel B. DV = Percentage of Students Receiving In-School Suspension				
Obs Received (25s)	0.31***	0.35**	-1.26**	-1.67*
	(0.07)	(0.13)	(0.45)	(0.83)
First Stage				
Obs Assigned (25s)			0.24***	0.18***
			(0.04)	(0.04)
F-Stat			43.06***	
Panel C. DV = Percentage of Students Receiving Out-of-School Suspension				
Obs Received (25s)	0.07	0.02	-0.73**	-1.41**
	(0.04)	(0.05)	(0.26)	(0.49)
First Stage				
Obs Assigned (25s)			0.24***	0.18***
			(0.04)	(0.04)
F-Stat			41.70***	
Panel D. DV = Percentage of Students Expelled				
Obs Received (25s)	0.01*	0.01	0.02	-0.01
	(0.00)	(0.00)	(0.02)	(0.03)
First Stage				
Obs Assigned (25s)			0.24***	0.18***
			(0.04)	(0.04)
F-Stat			41.65***	
School FE		X		X
2SLS			X	X
N(School-by-Years)	6329	6329	6329	6329

Notes: School-by-years are the unit of analysis. The outcomes are percentages ranging from zero to 100. The number of observations received and assigned are scaled by 25s; thus, coefficients represent association with an increase of 25 observations. Each model controls for year fixed effects and the prior-year 'outcome;' average teacher and administrator prior-year composite effectiveness scores; the proportions of students, teachers, and administrators who are white, black, and female; the proportions of students who are ELL, SPED, and economically disadvantaged; average teacher and administrator years of experience and level of education; the number of teachers in each school; and the ratio of teachers to evaluators in each school. Standard errors are clustered at the school level. Columns III and IV apply two-stage least-squares and report the first-stage coefficients. *F* statistics vary across panels in column III because each regression uses a different prior-year ‘outcome.’ Column IV applies the two-stage least-squares within estimator using Stata’s *xtivreg* command, which does not report a first-stage *F* statistic. * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$.

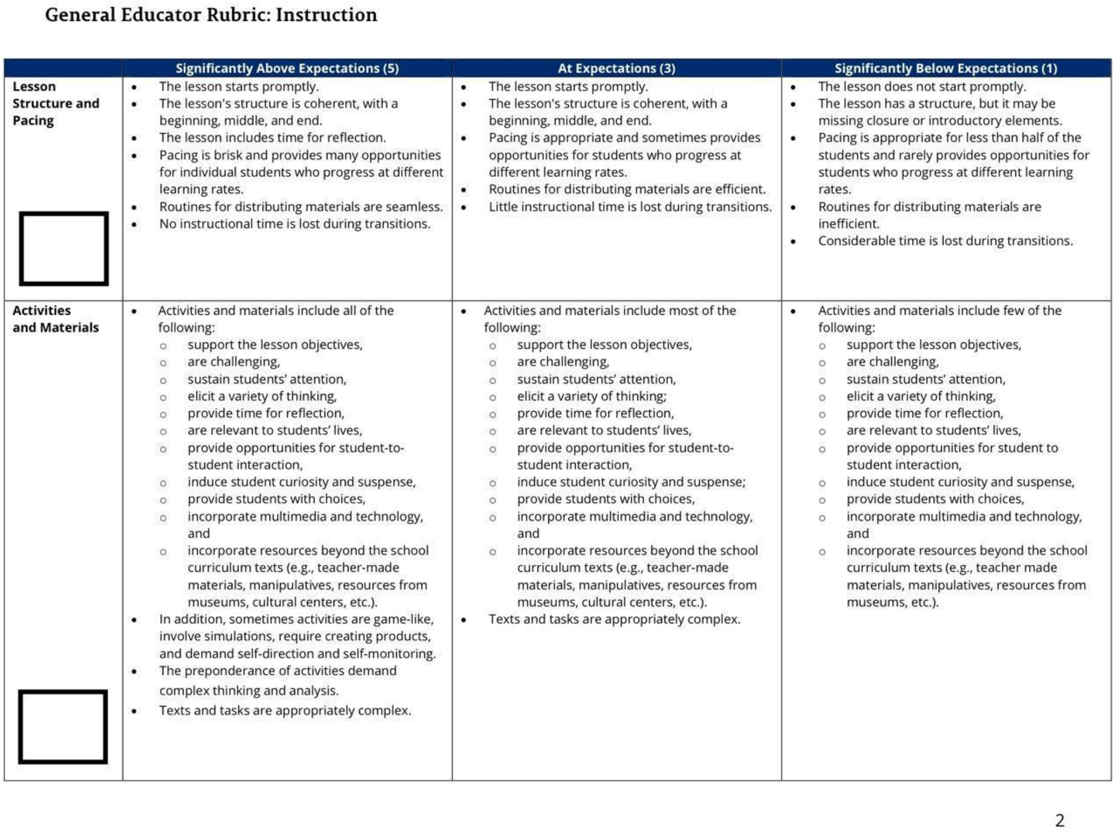
Online Appendix A. Standards-Based Performance Rubric



General Educator Rubric: Instruction

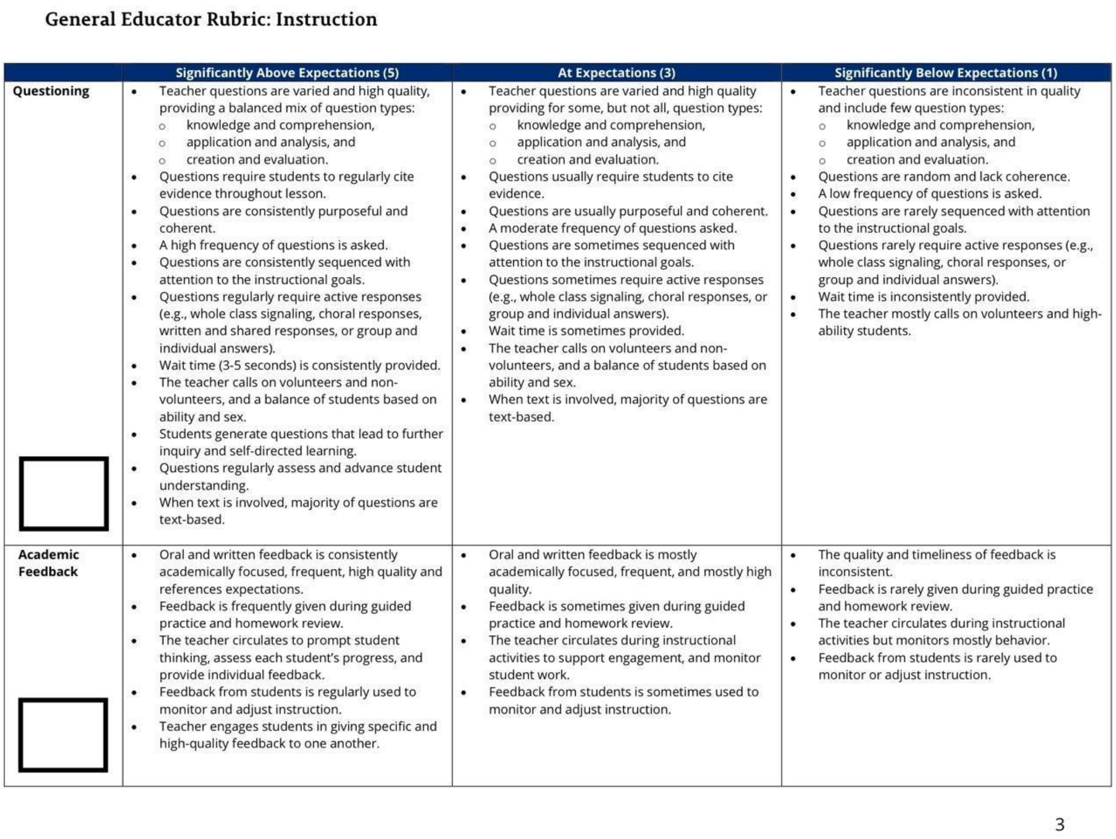
	Significantly Above Expectations (5)	At Expectations (3)	Significantly Below Expectations (1)
Standards and Objectives □	• All learning objectives are clearly and explicitly communicated, connected to state standards, and referenced throughout lesson. • Sub-objectives are aligned and logically sequenced to the lesson's major objective. • Learning objectives are: (a) consistently connected to what students have previously learned, (b) known from life experiences, and (c) integrated with other disciplines. • Expectations for student performance are clear, demanding, and high. • There is evidence that most students demonstrate mastery of the daily objective that supports significant progress towards mastery of a standard.	• Most learning objectives are communicated, connected to state standards, and referenced throughout lesson. • Sub-objectives are mostly aligned to the lesson's major objective. • Learning objectives are connected to what students have previously learned. • Expectations for student performance are clear. • There is evidence that most students demonstrate mastery of the daily objective that supports significant progress towards mastery of a standard.	• Few learning objectives are communicated, connected to state standards, and referenced throughout lesson. • Sub-objectives are inconsistently aligned to the lesson's major objective. • Learning objectives are rarely connected to what students have previously learned. • Expectations for student performance are vague. • There is evidence that few students demonstrate mastery of the daily objective that supports significant progress towards mastery of a standard.
Motivating Students □	• The teacher consistently organizes the content so that it is personally meaningful and relevant to students. • The teacher consistently develops learning experiences where inquiry, curiosity, and exploration are valued. • The teacher regularly reinforces and rewards effort.	• The teacher sometimes organizes the content so that it is personally meaningful and relevant to students. • The teacher sometimes develops learning experiences where inquiry, curiosity, and exploration are valued. • The teacher sometimes reinforces and rewards effort.	• The teacher rarely organizes the content so that it is personally meaningful and relevant to students. • The teacher rarely develops learning experiences where inquiry, curiosity, and exploration are valued. • The teacher rarely reinforces and rewards effort.
Presenting Instructional Content □	Presentation of content always includes: • visuals that establish the purpose of the lesson, preview the organization of the lesson, and include internal summaries of the lesson; • examples, illustrations, analogies, and labels for new concepts and ideas; • effective modeling of thinking process by the teacher and/or students guided by the teacher to demonstrate performance expectations; • concise communication; • logical sequencing and segmenting; • all essential information; and • no irrelevant, confusing, or non-essential information.	Presentation of content most of the time includes: • visuals that establish the purpose of the lesson, preview the organization of the lesson, and include internal summaries of the lesson; • examples, illustrations, analogies, and labels for new concepts and ideas; • modeling by the teacher to demonstrate performance expectations; • concise communication; • logical sequencing and segmenting; • all essential information; and • no irrelevant, confusing, or non-essential information.	Presentation of content rarely includes: • visuals that establish the purpose of the lesson, preview the organization of the lesson, and include internal summaries of the lesson; • examples, illustrations, analogies, and labels for new concepts and ideas; • modeling by the teacher to demonstrate performance expectations; • concise communication; • logical sequencing and segmenting; • all essential information; and • relevant, coherent, or essential information.

1



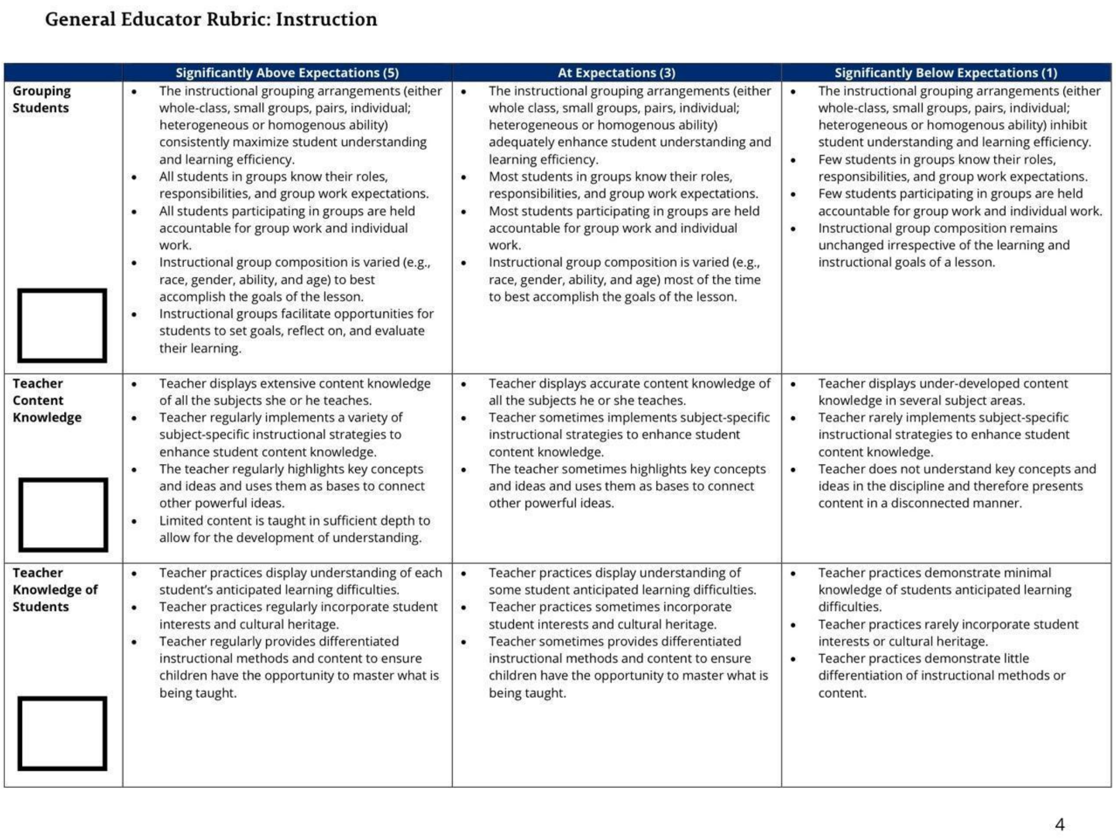
General Educator Rubric: Instruction

	Significantly Above Expectations (5)	At Expectations (3)	Significantly Below Expectations (1)
Lesson Structure and Pacing □	• The lesson starts promptly. • The lesson's structure is coherent, with a beginning, middle, and end. • The lesson includes time for reflection. • Pacing is brisk and provides many opportunities for individual students who progress at different learning rates. • Routines for distributing materials are seamless. • No instructional time is lost during transitions.	• The lesson starts promptly. • The lesson's structure is coherent, with a beginning, middle, and end. • Pacing is appropriate and sometimes provides opportunities for students who progress at different learning rates. • Routines for distributing materials are efficient. • Little instructional time is lost during transitions.	• The lesson does not start promptly. • The lesson has a structure, but it may be missing closure or introductory elements. • Pacing is appropriate for less than half of the students and rarely provides opportunities for students who progress at different learning rates. • Routines for distributing materials are inefficient. • Considerable time is lost during transitions.
Activities and Materials □	• Activities and materials include all of the following: ○ support the lesson objectives, ○ are challenging, ○ sustain students' attention, ○ elicit a variety of thinking, ○ provide time for reflection, ○ are relevant to students' lives, ○ provide opportunities for student-to-student interaction, ○ induce student curiosity and suspense, ○ provide students with choices, ○ incorporate multimedia and technology, and ○ incorporate resources beyond the school curriculum texts (e.g., teacher-made materials, manipulatives, resources from museums, cultural centers, etc.). • In addition, sometimes activities are game-like, involve simulations, require creating products, and demand self-direction and self-monitoring. • The preponderance of activities demand complex thinking and analysis. • Texts and tasks are appropriately complex.	• Activities and materials include most of the following: ○ support the lesson objectives, ○ are challenging, ○ sustain students' attention, ○ elicit a variety of thinking; ○ provide time for reflection, ○ are relevant to students' lives, ○ provide opportunities for student-to-student interaction, ○ induce student curiosity and suspense; ○ provide students with choices, ○ incorporate multimedia and technology, and ○ incorporate resources beyond the school curriculum texts (e.g., teacher-made materials, manipulatives, resources from museums, cultural centers, etc.). • Texts and tasks are appropriately complex.	• Activities and materials include few of the following: ○ support the lesson objectives, ○ are challenging, ○ sustain students' attention, ○ elicit a variety of thinking, ○ provide time for reflection, ○ are relevant to students' lives, ○ provide opportunities for student to student interaction, ○ induce student curiosity and suspense, ○ provide students with choices, ○ incorporate multimedia and technology, and ○ incorporate resources beyond the school curriculum texts (e.g., teacher made materials, manipulatives, resources from museums, etc.).



General Educator Rubric: Instruction

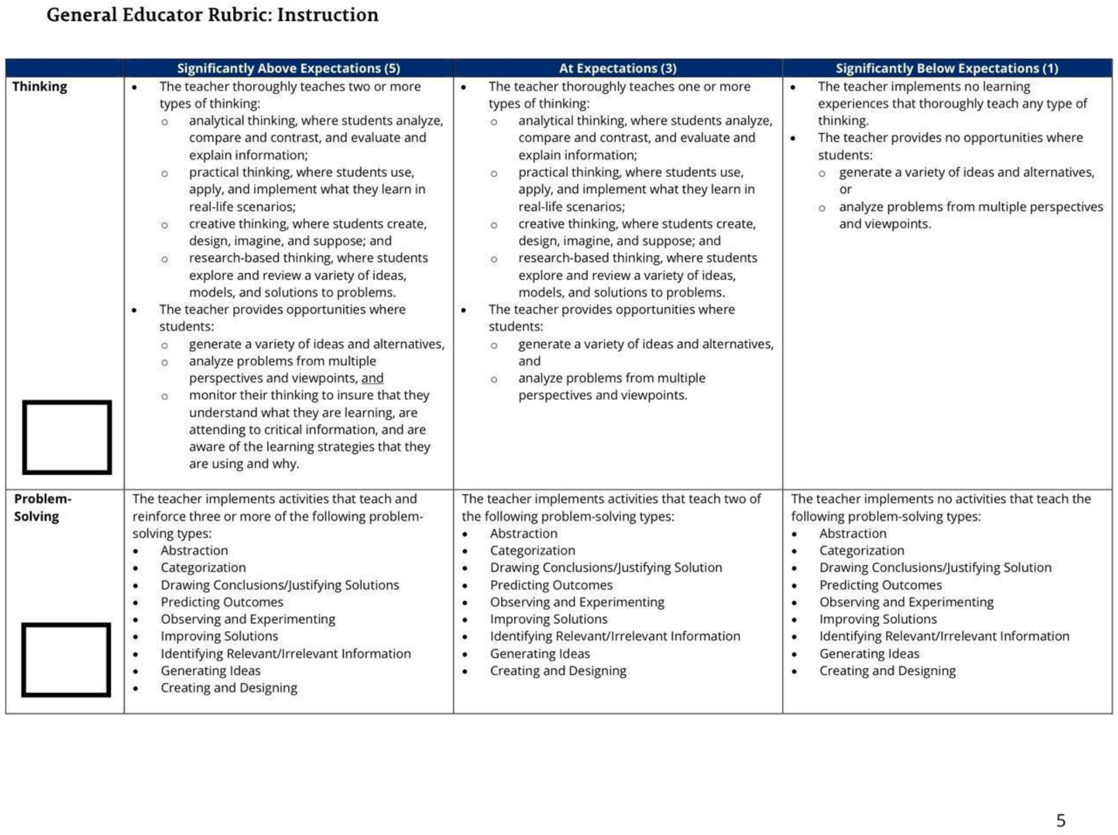
	Significantly Above Expectations (5)	At Expectations (3)	Significantly Below Expectations (1)
Questioning □	• Teacher questions are varied and high quality, providing a balanced mix of question types: ○ knowledge and comprehension, ○ application and analysis, and ○ creation and evaluation. • Questions require students to regularly cite evidence throughout lesson. • Questions are consistently purposeful and coherent. • A high frequency of questions is asked. • Questions are consistently sequenced with attention to the instructional goals. • Questions regularly require active responses (e.g., whole class signaling, choral responses, written and shared responses, or group and individual answers). • Wait time (3-5 seconds) is consistently provided. • The teacher calls on volunteers and non-volunteers, and a balance of students based on ability and sex. • Students generate questions that lead to further inquiry and self-directed learning. • Questions regularly assess and advance student understanding. • When text is involved, majority of questions are text-based.	• Teacher questions are varied and high quality providing for some, but not all, question types: ○ knowledge and comprehension, ○ application and analysis, and ○ creation and evaluation. • Questions usually require students to cite evidence. • Questions are usually purposeful and coherent. • A moderate frequency of questions asked. • Questions are sometimes sequenced with attention to the instructional goals. • Questions sometimes require active responses (e.g., whole class signaling, choral responses, or group and individual answers). • Wait time is sometimes provided. • The teacher calls on volunteers and non-volunteers, and a balance of students based on ability and sex. • When text is involved, majority of questions are text-based.	• Teacher questions are inconsistent in quality and include few question types: ○ knowledge and comprehension, ○ application and analysis, and ○ creation and evaluation. • Questions are random and lack coherence. • A low frequency of questions is asked. • Questions are rarely sequenced with attention to the instructional goals. • Questions rarely require active responses (e.g., whole class signaling, choral responses, or group and individual answers). • Wait time is inconsistently provided. • The teacher mostly calls on volunteers and high-ability students.
Academic Feedback □	• Oral and written feedback is consistently academically focused, frequent, high quality and references expectations. • Feedback is frequently given during guided practice and homework review. • The teacher circulates to prompt student thinking, assess each student's progress, and provide individual feedback. • Feedback from students is regularly used to monitor and adjust instruction. • Teacher engages students in giving specific and high-quality feedback to one another.	• Oral and written feedback is mostly academically focused, frequent, and mostly high quality. • Feedback is sometimes given during guided practice and homework review. • The teacher circulates during instructional activities to support engagement, and monitor student work. • Feedback from students is sometimes used to monitor and adjust instruction.	• The quality and timeliness of feedback is inconsistent. • Feedback is rarely given during guided practice and homework review. • The teacher circulates during instructional activities but monitors mostly behavior. • Feedback from students is rarely used to monitor or adjust instruction.



General Educator Rubric: Instruction

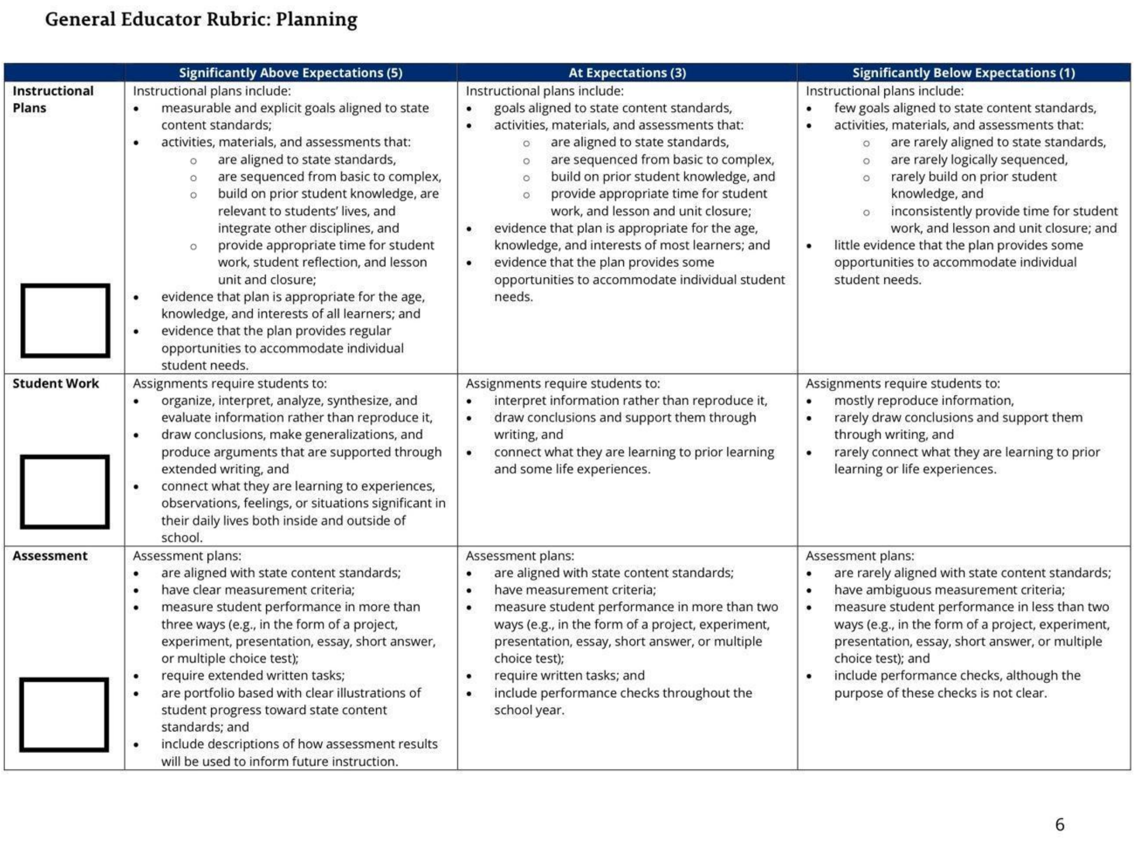
	Significantly Above Expectations (5)	At Expectations (3)	Significantly Below Expectations (1)
Grouping Students □	• The instructional grouping arrangements (either whole-class, small groups, pairs, individual; heterogeneous or homogenous ability) consistently maximize student understanding and learning efficiency. • All students in groups know their roles, responsibilities, and group work expectations. • All students participating in groups are held accountable for group work and individual work. • Instructional group composition is varied (e.g., race, gender, ability, and age) to best accomplish the goals of the lesson. • Instructional groups facilitate opportunities for students to set goals, reflect on, and evaluate their learning.	• The instructional grouping arrangements (either whole class, small groups, pairs, individual; heterogeneous or homogenous ability) adequately enhance student understanding and learning efficiency. • Most students in groups know their roles, responsibilities, and group work expectations. • Most students participating in groups are held accountable for group work and individual work. • Instructional group composition is varied (e.g., race, gender, ability, and age) most of the time to best accomplish the goals of the lesson.	• The instructional grouping arrangements (either whole-class, small groups, pairs, individual; heterogeneous or homogenous ability) inhibit student understanding and learning efficiency. • Few students in groups know their roles, responsibilities, and group work expectations. • Few students participating in groups are held accountable for group work and individual work. • Instructional group composition remains unchanged irrespective of the learning and instructional goals of a lesson.
Teacher Content Knowledge □	• Teacher displays extensive content knowledge of all the subjects she or he teaches. • Teacher regularly implements a variety of subject-specific instructional strategies to enhance student content knowledge. • The teacher regularly highlights key concepts and ideas and uses them as bases to connect other powerful ideas. • Limited content is taught in sufficient depth to allow for the development of understanding.	• Teacher displays accurate content knowledge of all the subjects he or she teaches. • Teacher sometimes implements subject-specific instructional strategies to enhance student content knowledge. • The teacher sometimes highlights key concepts and ideas and uses them as bases to connect other powerful ideas.	• Teacher displays under-developed content knowledge in several subject areas. • Teacher rarely implements subject-specific instructional strategies to enhance student content knowledge. • Teacher does not understand key concepts and ideas in the discipline and therefore presents content in a disconnected manner.
Teacher Knowledge of Students □	• Teacher practices display understanding of each student's anticipated learning difficulties. • Teacher practices regularly incorporate student interests and cultural heritage. • Teacher regularly provides differentiated instructional methods and content to ensure children have the opportunity to master what is being taught.	• Teacher practices display understanding of some student anticipated learning difficulties. • Teacher practices sometimes incorporate student interests and cultural heritage. • Teacher sometimes provides differentiated instructional methods and content to ensure children have the opportunity to master what is being taught.	• Teacher practices demonstrate minimal knowledge of students anticipated learning difficulties. • Teacher practices rarely incorporate student interests or cultural heritage. • Teacher practices demonstrate little differentiation of instructional methods or content.

4



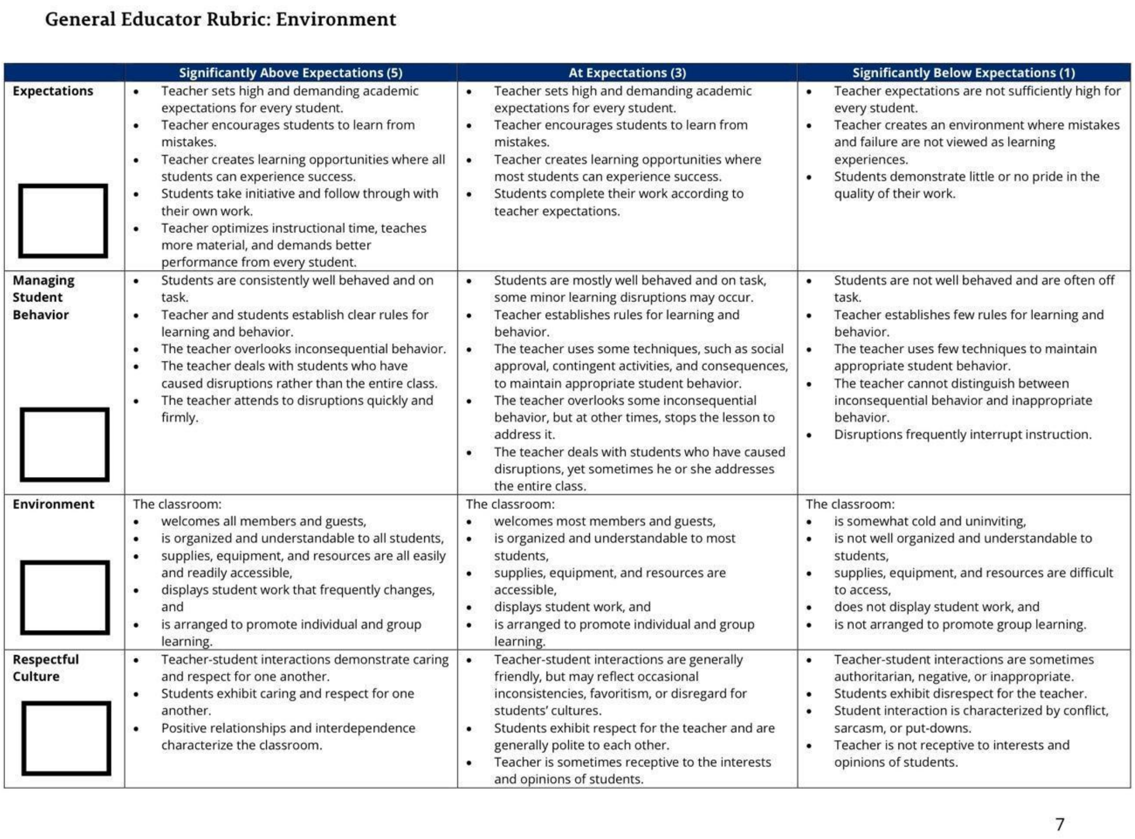
General Educator Rubric: Instruction

	Significantly Above Expectations (5)	At Expectations (3)	Significantly Below Expectations (1)
Thinking □	- The teacher thoroughly teaches two or more types of thinking: ○ analytical thinking, where students analyze, compare and contrast, and evaluate and explain information; ○ practical thinking, where students use, apply, and implement what they learn in real-life scenarios; ○ creative thinking, where students create, design, imagine, and suppose; and ○ research-based thinking, where students explore and review a variety of ideas, models, and solutions to problems. - The teacher provides opportunities where students: ○ generate a variety of ideas and alternatives, ○ analyze problems from multiple perspectives and viewpoints, and ○ monitor their thinking to insure that they understand what they are learning, are attending to critical information, and are aware of the learning strategies that they are using and why.	- The teacher thoroughly teaches one or more types of thinking: ○ analytical thinking, where students analyze, compare and contrast, and evaluate and explain information; ○ practical thinking, where students use, apply, and implement what they learn in real-life scenarios; ○ creative thinking, where students create, design, imagine, and suppose; and ○ research-based thinking, where students explore and review a variety of ideas, models, and solutions to problems. - The teacher provides opportunities where students: ○ generate a variety of ideas and alternatives, and ○ analyze problems from multiple perspectives and viewpoints.	- The teacher implements no learning experiences that thoroughly teach any type of thinking. - The teacher provides no opportunities where students: ○ generate a variety of ideas and alternatives, or ○ analyze problems from multiple perspectives and viewpoints.
Problem-Solving □	The teacher implements activities that teach and reinforce three or more of the following problem-solving types: - Abstraction - Categorization - Drawing Conclusions/Justifying Solutions - Predicting Outcomes - Observing and Experimenting - Improving Solutions - Identifying Relevant/Irrelevant Information - Generating Ideas - Creating and Designing	The teacher implements activities that teach two of the following problem-solving types: - Abstraction - Categorization - Drawing Conclusions/Justifying Solution - Predicting Outcomes - Observing and Experimenting - Improving Solutions - Identifying Relevant/Irrelevant Information - Generating Ideas - Creating and Designing	The teacher implements no activities that teach the following problem-solving types: - Abstraction - Categorization - Drawing Conclusions/Justifying Solution - Predicting Outcomes - Observing and Experimenting - Improving Solutions - Identifying Relevant/Irrelevant Information - Generating Ideas - Creating and Designing



General Educator Rubric: Planning

	Significantly Above Expectations (5)	At Expectations (3)	Significantly Below Expectations (1)
Instructional Plans □	Instructional plans include: • measurable and explicit goals aligned to state content standards; • activities, materials, and assessments that: o are aligned to state standards, o are sequenced from basic to complex, o build on prior student knowledge, are relevant to students' lives, and integrate other disciplines, and o provide appropriate time for student work, student reflection, and lesson unit and closure; • evidence that plan is appropriate for the age, knowledge, and interests of all learners; and • evidence that the plan provides regular opportunities to accommodate individual student needs.	Instructional plans include: • goals aligned to state content standards, • activities, materials, and assessments that: o are aligned to state standards, o are sequenced from basic to complex, o build on prior student knowledge, and o provide appropriate time for student work, and lesson and unit closure; • evidence that plan is appropriate for the age, knowledge, and interests of most learners; and • evidence that the plan provides some opportunities to accommodate individual student needs.	Instructional plans include: • few goals aligned to state content standards, • activities, materials, and assessments that: o are rarely aligned to state standards, o are rarely logically sequenced, o rarely build on prior student knowledge, and o inconsistently provide time for student work, and lesson and unit closure; and • little evidence that the plan provides some opportunities to accommodate individual student needs.
Student Work □	Assignments require students to: • organize, interpret, analyze, synthesize, and evaluate information rather than reproduce it, • draw conclusions, make generalizations, and produce arguments that are supported through extended writing, and • connect what they are learning to experiences, observations, feelings, or situations significant in their daily lives both inside and outside of school.	Assignments require students to: • interpret information rather than reproduce it, • draw conclusions and support them through writing, and • connect what they are learning to prior learning and some life experiences.	Assignments require students to: • mostly reproduce information, • rarely draw conclusions and support them through writing, and • rarely connect what they are learning to prior learning or life experiences.
Assessment □	Assessment plans: • are aligned with state content standards; • have clear measurement criteria; • measure student performance in more than three ways (e.g., in the form of a project, experiment, presentation, essay, short answer, or multiple choice test); • require extended written tasks; • are portfolio based with clear illustrations of student progress toward state content standards; and • include descriptions of how assessment results will be used to inform future instruction.	Assessment plans: • are aligned with state content standards; • have measurement criteria; • measure student performance in more than two ways (e.g., in the form of a project, experiment, presentation, essay, short answer, or multiple choice test); • require written tasks; and • include performance checks throughout the school year.	Assessment plans: • are rarely aligned with state content standards; • have ambiguous measurement criteria; • measure student performance in less than two ways (e.g., in the form of a project, experiment, presentation, essay, short answer, or multiple choice test); and • include performance checks, although the purpose of these checks is not clear.



General Educator Rubric: Environment

	Significantly Above Expectations (5)	At Expectations (3)	Significantly Below Expectations (1)
Expectations □	• Teacher sets high and demanding academic expectations for every student. • Teacher encourages students to learn from mistakes. • Teacher creates learning opportunities where all students can experience success. • Students take initiative and follow through with their own work. • Teacher optimizes instructional time, teaches more material, and demands better performance from every student.	• Teacher sets high and demanding academic expectations for every student. • Teacher encourages students to learn from mistakes. • Teacher creates learning opportunities where most students can experience success. • Students complete their work according to teacher expectations.	• Teacher expectations are not sufficiently high for every student. • Teacher creates an environment where mistakes and failure are not viewed as learning experiences. • Students demonstrate little or no pride in the quality of their work.
Managing Student Behavior □	• Students are consistently well behaved and on task. • Teacher and students establish clear rules for learning and behavior. • The teacher overlooks inconsequential behavior. • The teacher deals with students who have caused disruptions rather than the entire class. • The teacher attends to disruptions quickly and firmly.	• Students are mostly well behaved and on task, some minor learning disruptions may occur. • Teacher establishes rules for learning and behavior. • The teacher uses some techniques, such as social approval, contingent activities, and consequences, to maintain appropriate student behavior. • The teacher overlooks some inconsequential behavior, but at other times, stops the lesson to address it. • The teacher deals with students who have caused disruptions, yet sometimes he or she addresses the entire class.	• Students are not well behaved and are often off task. • Teacher establishes few rules for learning and behavior. • The teacher uses few techniques to maintain appropriate student behavior. • The teacher cannot distinguish between inconsequential behavior and inappropriate behavior. • Disruptions frequently interrupt instruction.
Environment □	The classroom: • welcomes all members and guests, • is organized and understandable to all students, • supplies, equipment, and resources are all easily and readily accessible, • displays student work that frequently changes, and • is arranged to promote individual and group learning.	The classroom: • welcomes most members and guests, • is organized and understandable to most students, • supplies, equipment, and resources are accessible, • displays student work, and • is arranged to promote individual and group learning.	The classroom: • is somewhat cold and uninviting, • is not well organized and understandable to students, • supplies, equipment, and resources are difficult to access, • does not display student work, and • is not arranged to promote group learning.
Respectful Culture □	• Teacher-student interactions demonstrate caring and respect for one another. • Students exhibit caring and respect for one another. • Positive relationships and interdependence characterize the classroom.	• Teacher-student interactions are generally friendly, but may reflect occasional inconsistencies, favoritism, or disregard for students' cultures. • Students exhibit respect for the teacher and are generally polite to each other. • Teacher is sometimes receptive to the interests and opinions of students.	• Teacher-student interactions are sometimes authoritarian, negative, or inappropriate. • Students exhibit disrespect for the teacher. • Student interaction is characterized by conflict, sarcasm, or put-downs. • Teacher is not receptive to interests and opinions of students.

Online Appendix B. Sensitivity and Falsification Analyses

Local Regressions

We identify teachers in the bandwidths of 10, 20, and 30 surrounding each cutoff that assigns observations, find the number of observations assigned to these teachers, then find each school-by-year total. We then apply Equation B1:

$$y_{st} = \delta ass\widetilde{igned}_{bst} + X_{st} + \alpha_t + \eta_{bst} + e_{st} \, , \qquad \text{(B1)}$$

where $ass\widetilde{igned}_{bst}$ is the number of observations assigned to the teachers in bandwidth *b* surrounding the prior-year continuous LOE thresholds of 200 and 425, η_{bst} is the proportion of teachers from school *s* in year *t* who contributed to $ass\widetilde{igned}_{bst}$ and whose prior-year continuous LOE came from the bandwidth surrounding the 425 threshold. All other terms are identical to those from previous equations; that is, they are based on all student, teachers, and administrators in school *s* in year *t*, not just the teachers in bandwidth *b*. We characterize the regression of SDOs linked to all teachers' students in school *s* in year *t* on the number of observations assigned to the teachers in bandwidth *b* as introducing measurement error because y_{st} presumably overstates the number of SDOs linked to teachers in *b*, which likely attenuates our estimates. Standard errors are clustered at the school level.

Table B1

Sensitivity Tests: Local Regressions

	I	II	III
	w = 10	*w* = 20	*w* = 30
Panel A. DV = Percentage of Students Receiving Any Suspension/ Expulsion			
Obs Assigned (25s)	-0.16	-0.11	-0.10
	(0.21)	(0.13)	(0.10)
Panel B. DV = Percentage of Students Receiving In-School Suspension			
Obs Assigned (25s)	-0.08	-0.06	-0.08
	(0.23)	(0.14)	(0.11)
Panel C. DV = Percentage of Students Receiving Out-of-School Suspension			
Obs Assigned (25s)	-0.09	-0.06	-0.04
	(0.23)	(0.07)	(0.06)
Panel D. DV = Percentage of Students Expelled			
Obs Assigned (25s)	0.00	0.00	0.00
	(0.01)	(0.01)	(0.01)
N(School-by-Years)	6329	6329	6329

Notes: School-by-years are the unit of analysis. The outcomes are percentages ranging from zero to 100. The independent variable, the number of observations assigned, only uses observations assigned to teachers whose prior-year LOE was within [200 – *w*, 200 + *w*] or [425 – *w*, 425 + *w*]. All other variables use all teachers; thus, none of the other variables changed in these models. The independent variable is still scaled by 25s; thus, coefficients represent the change in an outcome associated with an increase of 25 policy-assigned observations. Each model controls for year fixed effects and the prior-year 'outcome;' average teacher and administrator prior-year composite effectiveness scores; the proportions of students, teachers, and administrators who are white, black, and female; the proportions of students who are ELL, SPED, and economically disadvantaged; average teacher and administrator years of experience and level of education; the number of teachers in each school; and the ratio of teachers to evaluators in each school. Standard errors are clustered at the school level. * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$.

Table B2

Rothstein Falsification Tests

	I	II	III	IV
	ISS/ OSS/ EX	ISS-only	OSS-only	EX-only
Obs Assigned (25s)	0.01	0.01	0.00	0.00
	(0.01)	(0.01)	(0.00)	(0.00)
N(School-by-Years)	5284	5284	5284	5284

Notes: School-by-years are the unit of analysis. The outcomes are percentages ranging from zero to 100. The independent variable, the number of *next-year* observations assigned, is scaled by 25s; thus, coefficients represent the change in an outcome associated with an increase of 25 policy-assigned observations *in the next year*. Each model controls for school fixed effects and year fixed effects and the prior-year 'outcome;' average teacher and administrator prior-year composite effectiveness scores; the proportions of students, teachers, and administrators who are white, black, and female; the proportions of students who are ELL, SPED, and economically disadvantaged; average teacher and administrator years of experience and level of education; the number of teachers in each school; and the ratio of teachers to evaluators in each school. Standard errors are clustered at the school level. * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$.

Online Appendix C. Survey Item Mediation Analyses

Table C1 Panel A prints mediation analysis results concerning the *Administrator Presence*. Panel A uses the sample of school-by-years with at least one TES response to the *Administrator Presence* survey item. Despite the substantial reduction in sample size and restriction to two years of data, the unmediated relationships in Table C1 Panel A Column I are qualitatively consistent with findings based on the full sample (see Table 2): assigning schools more observations is associated with reductions in the percentage of students receiving at least one ISS or one OSS. If *Administrator Presence* mediates these relationships as we hypothesized, we will detect positive relationships in Table C1 Panel A Column II; instead, the data suggest that assigning schools 25 more observations is associated with a statistically significant 0.05 unit (0.15 SD) decline in the *Administrator Presence* variable. However, Panel A Columns III and IV provide no evidence supporting our hypothesis that *Administrator Presence* is a mediator.

Table C1 Panel B reports the relationship between policy-assigned observations and SDOs using the sample of school-by-years with at least one TES response to the *Administrator Time* survey item. As we detect no total main relationship in the Panel B subsample (Column I), there is no main relationship to mediate.

The mediation analyses using survey items do not distinguish between school-by-years in which most potential survey respondents submitted a survey and those in which few or one submitted a survey. We incorporate school-by-year response rates into the analyses from Table C1 using weighted linear regressions, where the weight applied to school *s* in year *t* is the proportion of potential survey respondents in school *s* in year *t* who responded to the survey. The range of weights used in the *Administrator Presence* analyses (teacher survey item) ranges from 0.01 to 1.0, with a 0.33 mean and 0.16 standard deviation; the range of weights used in the *Administrator Time* (administrator survey item) ranges from 0.14 to 1.0, with a 0.75 mean and 0.28 standard deviation. While the weights affect some estimates (Table C2), they do not affect our inferences: no evidence suggests that the *Administrator Time* or *Presence* measures are mediators.

Table C1

Mediation by Administrator Survey Items Regarding Administrator Presence and Time

	I	II	III	IV
	Total Main Relationship	Mediator Relationship	Mediated Main Relationship	% change in main relationship
Panel A. Administrator Presence Mediator				
Panel A1. DV = In-School Suspensions				
Obs Assigned (25s)	-0.69*	-0.05*	-0.69*	-0%
	(-1.29, -0.09)	(-0.10, -0.002)	(-1.29, -0.09)	
Panel A2. DV = Out-of-School Suspensions				
Obs Assigned (25s)	-0.37**	-0.05*	-0.36**	+3%
	(-0.61, -0.13)	(-0.10, -0.002)	(-0.60, -0.12)	
N(School-by-Year)	2163	2163	2163	
Panel B. Administrator Time Mediator				
Panel B1. DV = In-School Suspensions				
Obs Assigned (25s)	-0.26	0.27	-0.25	-4%
	(-0.66, 0.15)	(-0.15, 0.68)	(-0.65, 0.16)	
Panel B2. DV = Out-of-School Suspensions				
Obs Assigned (25s)	-0.12	0.27	-0.12	0%
	(-0.26, 0.03)	(-0.15, 0.68)	(-0.26, 0.03)	
N(School-by-Year)	3042	3042	3042	

Notes: School-by-years are the unit of analysis. Standard errors are clustered at the school level; 95% confidence intervals in parentheses. The discipline outcomes are percentages ranging from zero to 100; in reverse-coded terms, the *Administrator Presence* item ranges from (1=Strongly Agree) to (4=Strongly Disagree), with a 2.62 mean and 0.34 standard deviation; the *Administrator Time* item ranges from 0 to 51, with a 6.42 mean and 5.44 standard deviation. The number of observations assigned are scaled by 25s; thus, coefficients represent association with an increase of 25 observations. Column I lists unmediated results using the same sample as the mediation analysis. Column II treats the mediator as the 'outcome.' Column III uses the same outcome and right hand-side variables as column I but adds the mediator as a right hand-side variable. Each model controls for year and school fixed effects and the prior-year 'outcome;' average teacher and administrator prior-year composite effectiveness scores; the proportions of students, teachers, and administrators who are white, black, and female; the proportions of students who are ELL, SPED, and economically disadvantaged; average teacher and administrator years of experience and level of education; the number of teachers in each school; and the ratio of teachers to evaluators in each school.

Table C2

Original and Mediation Regressions Weighted by the Proportion of School-by-Year Survey Respondents

	I	II	III	IV
	Total Main Relationship	Mediator Relationship	Mediated Main Relationship	% change in main relationship
Panel A. Administrator Presence Mediator				
Panel A1. DV = In-School Suspensions				
Obs Assigned (25s)	-0.43	-0.03	-0.45	+5%
	(-1.08, 0.22)	(-0.07, 0.01)	(-1.10, 0.20)	
Panel A2. DV = Out-of-School Suspensions				
Obs Assigned (25s)	-0.30*	-0.03	-0.29*	-3%
	(-0.55, -0.04)	(-0.07, 0.01)	(-0.55, -0.04)	
N(School-by-Year)	2163	2163	2163	
Panel B. Administrator Time Mediator				
Panel B1. DV = In-School Suspensions				
Obs Assigned (25s)	-0.35	0.30	-0.35	0%
	(-0.77, 0.06)	(-0.08, 0.68)	(-0.77, 0.07)	
Panel B2. DV = Out-of-School Suspensions				
Obs Assigned (25s)	-0.15*	0.30	-0.15*	0%
	(-0.30, -0.002)	(-0.08, 0.68)	(-0.31, -0.004)	
N(School-by-Year)	3042	3042	3042	

Notes: School-by-years are the unit of analysis. Standard errors are clustered at the school level; 95% confidence intervals in parentheses. Results in every column generated by weighted linear regressions, where the weight of school *s* in year *t* is the proportion of potential survey respondents in school *s* in year *t* who responded to the survey in year *t*. The range of weights used in Panel A (teacher survey item) ranges from 0.01 to 1.0, with a 0.33 mean and 0.16 standard deviation; the range of weights used in Panel B (administrator survey item) ranges from 0.14 to 1.0, with a 0.75 mean and 0.28 standard deviation. The discipline outcomes are percentages ranging from zero to 100; in reverse-coded terms, the teacher survey item from Panel A ranges from (1=Strongly Agree) to (4=Strongly Disagree), with a 2.62 mean and 0.34 standard deviation; the administrator survey item in Panel B ranges from 0 to 51, with a 6.42 mean and 5.44 standard deviation. The number of observations assigned are scaled by 25s; thus, coefficients represent association with an increase of 25 observations. Column I lists unmediated results using the same sample as the mediation analyses. Column II treats the mediator as the ‘outcome.’ Column III uses the same outcome and right hand-side variables as column I but adds the mediator as a right hand-side variable. Each model controls for year and school fixed effects and the prior-year 'outcome;' average teacher and administrator prior-year composite effectiveness scores; the proportions of students, teachers, and administrators who are white, black, and female; the proportions of students who are ELL, SPED, and economically disadvantaged; average teacher and administrator years of experience and level of education; the number of teachers in each school; and the ratio of teachers to evaluators in each school.

Online Appendix D. In-School Suspensions: Moderation Analyses

Table D1

Moderated Relationships Between Policy-Assigned Observations and ISS

	I	II	III
Obs Assigned (25s)	-0.15	0.37	-0.09
	(0.15)	(0.48)	(0.45)
Obs Assigned (25s) * Prior-Year OSS	-0.01		
	(0.01)		
Obs Assigned (25s) * Avg Tch Prior-Year LOE (50s)		-0.09	
		(0.06)	
Obs Assigned (25s) * Avg Tch Yrs Experience			-0.02
			(0.04)
N(School-by-Years)	6329	6329	6329

Notes: School-by-years are the unit of analysis. Each column-by-panel presents results from a separate regression. The outcomes are percentages ranging from zero to 100. The independent variable, the number of observations assigned, is scaled by 25s; thus, coefficients represent the change in an outcome associated with an increase of 25 policy-assigned observations. The average teacher years of experience and prior-year ISS are scaled by ones; average teacher prior-year LOE is scaled by 50. Each model includes year and school fixed effects and controls for the prior-year 'outcome;' average teacher and administrator prior-year composite effectiveness scores; the proportions of students, teachers, and administrators who are white, black, and female; the proportions of students who are ELL, SPED, and economically disadvantaged; average teacher and administrator years of experience and level of education; the number of teachers in each school; and the ratio of teachers to evaluators in each school. Standard errors are clustered at the school level.