education policy analysis archives

A peer-reviewed, independent,
open access, multilingual journal

epaa aape

Arizona State University

Volume 26 Number 41 April 2, 2018 ISSN 1068-2341

Teacher Perceptions of a New Performance Evaluation System and Their Influence on Practice: A Within- and Between-School Level Analysis

Matthew Finster
&
Anthony Milanowski
Westat
United States

Citation: Finster, M., & Milanowski, A. (2018). Teacher perceptions of a new performance evaluation system and their influence on practice: A within- and between-school level analysis. *Education Policy Analysis Archives, 26*(41). http://dx.doi.org/10.14507/epaa.26.3500

Abstract: Teacher performance evaluation systems (PESs) are central to policy efforts to increase teacher effectiveness and student learning. We argue that for these reforms to work, PESs need to be treated as coherent systems, in which teachers perceive that there are linkages between the PES components. Using teacher survey data from a large, midwestern school district, this article explores the linkages between teacher perceptions of a new PES using confirmatory factor analysis (CFA), structural equation modeling (SEM), and multilevel CFA. We also examine whether a strong evaluation climate developed in this district. The CFA and SEM analysis demonstrate that teacher perceptions of PES are interrelated and linked to perceptions of changes in teaching practices and to the potential impact on student learning. The multilevel CFA demonstrates cross-level noninvariance, with fewer factors being identified at the school levels. These results suggest a need for a school-level theory of action with corresponding school-level constructs. While we did not find evidence of a shared strong evaluation climate, the results of the analysis illustrate the importance of examining within-school agreement, both to assess the reliability of between-school differences in average teacher perceptions and to assess whether schools are developing a strong evaluation climate.

Journal website: http://epaa.asu.edu/ojs/
Facebook: /EPAAA
Twitter: @epaa_aape

Manuscript received: 10/27/2017
Revisions received: 1/10/2018
Accepted: 2/14/2018

Keywords: policy; teacher evaluation system; teacher perceptions; school climate; structural equation modeling

Percepciones de los maestros sobre un nuevo sistema de evaluación del desempeño y su influencia en la práctica: Un análisis dentro de las escuelas y entre las escuelas

Resumen: Los sistemas de evaluación del desempeño de los maestros (PES) son fundamentales para los esfuerzos de las políticas para aumentar la efectividad de los docentes y el aprendizaje de los estudiantes. Argumentamos que para que estas reformas funcionen, los SPE deben tratarse como sistemas coherentes, en los que los docentes perciben que existen vínculos entre los componentes de PSA. Utilizando los datos de la encuesta de docentes de un gran distrito escolar del medio oeste, este artículo explora los vínculos entre las percepciones de los docentes de un nuevo PES utilizando análisis factorial confirmatorio (CFA), modelado de ecuaciones estructurales (SEM) y CFA multinivel. También examinamos si se desarrolló un clima de evaluación fuerte en este distrito. Los análisis de CFA y SEM demuestran que las percepciones de los maestros sobre los SPE se interrelacionan y se relacionan con las percepciones de los cambios en las prácticas de enseñanza y con el posible impacto en el aprendizaje de los estudiantes. El CFA multinivel demuestra la no invariabilidad de niveles cruzados, con menos factores identificados en los niveles escolares. Estos resultados sugieren la necesidad de una teoría de acción a nivel escolar con los constructos correspondientes a nivel escolar. Aunque no encontramos evidencia de un clima de evaluación sólido compartido, los resultados del análisis ilustran la importancia de examinar el acuerdo dentro de la escuela, tanto para evaluar la confiabilidad de las diferencias entre las escuelas en las percepciones promedio de los docentes como para evaluar si las escuelas están desarrollando clima de evaluación fuerte.
Palabras clave: política; sistema de evaluación de los maestros; percepciones del maestro; clima escolar; modelos de ecuaciones estructurales

Percepções dos maestros sobre um novo sistema de avaliação da performance e da sua influência na prática: Uma análise dentro das escuelas e entre as escuelas

Resumo: Los sistemas de avaliação do desempenho dos professores (PES) filho fundamental para os esforços das políticas para aumentar a eficácia dos professores e a aprendizagem dos estudiosos. Argumentamos que, para que estes esquemas funcionem, os SPE deben tratarse como os sistemas coerentes, nos quais os docentes percebem que existem entre os componentes da PSA. Utilizando os dados da pesquisa de professores de um grande distrito da medicina ocidental, este artigo explora os acontecimentos entre as percepções dos docentes de um novo PES analisado factorial confirmatorio (CFA), modelo de ecuaciones estructurales (SEM) e CFA multinivel. Também foi possível encontrar um clima de avaliação para avaliar este distrito. As análises de CFA e SEM demonstram quais são as percepções das mudanças nas relações públicas e se relacionam com a percepção das mudanças nas prácticas de enseñanza e com o possível impacto na aprendizagem dos estudiosos. El CFA multinivel demuestra a invariabilidade de niveles cruzados, com menos fatores identificados nas niveles escolares. Estos resultados sugam a necessidade de uma teoría de uma escola nivelada com os constructos correspondentes a nivel escolar. A inexistente evidencias in the clima of evaluación maciço compartimenta, los procesos del análisis ilustrant la importancia de dashboard in aeroplane of la escuela, both for evaluate the dependibility of las différences en las escuelas en las percepción promedio de los docentes for evaluat si as escuelas estão desarrollando clima de avaliação fuerte.
Palavras-chave: política; sistema de avaliação dos maestros; percepciones del maestro; clima escolar; modelos de ecuaciones estructurales

Introduction

Recent federal and state initiatives have encouraged districts to overhaul teacher evaluation systems to include multiple measures and to provide move valid information to differentiate decisions about educator human capital management.[1] Today, most states have designed and adopted new teacher evaluation systems (Steinberg & Donaldson, 2016). While scholars argue over the primary role of performance evaluation systems (PESs), theories of action point to two main pathways that may increase student achievement. One is through improving the instructional practice of existing teachers through frequent instructional feedback based on classroom observation, setting goals for improved instruction, and providing relevant professional development opportunities (e.g., Papay, 2012). The second, acting over a longer term, uses evaluation results to identify higher performing teachers for recognition and rewards, improving their retention, and signaling lower performing teachers to improve, consider changing careers, or be terminated (e.g., Hanushek, 2009).

Our theory of action draws on research (e.g., Hedge & Teachout, 2000; Milanowski & Heneman, 2001; O'Pry & Schumacher, 2012) that suggests that evaluatees' perceptions of the evaluation systems affect whether these systems will have the positive impacts on performance desired and whether they will be sustained. These findings are consistent with other research on educational innovations that highlights the importance of considering teachers' perceptions in response to reforms designed to change practice (e.g., Datnow & Catellano, 2000; Gitlin & Margonis, 1995). In particular, teachers' understanding the new system, perceptions of evaluator credibility, the quality of feedback, and the fairness of the performance measures are likely to be related to whether teachers use the evaluation results to improve their practice and support continuing the system (Cherasaro, Brodersen, Reale, & Yanoski, 2016; Heneman & Milanowski, 2003; Williams & Levy, 2000).

Several studies have examined the relationships between teacher perceptions of evaluation and indicated that the factors are interrelated (e.g., Cherasaro et al., 2016; Jiang, Sporte, & Luppescu, 2015). Building on this literature, using survey data from a large, midwestern U.S. school district, this study uses confirmatory factor analysis (CFA) and structural equation modeling (SEM) to identify latent factors of teacher perceptions of evaluation and to examine the relationships between the multiple latent factors. This analysis provides evidence for the validity of the scales developed to assess teacher perceptions and confirms that expected relationships postulated in the theory of action linking teacher evaluation to improved performance are in place. We find that teacher perceptions of PES characteristics (e.g., fairness of measures, evaluator credibility, and quality of feedback) influence the perceived impact of the evaluation process on teaching and teacher perceptions of its potential benefits

We also assess the multilevel nature of teacher perceptions at the group level (school level). To date, we are unaware of any studies that have examined the multilevel factorial structure (i.e., measurement noninvariance or cross-level invariance) of teacher perceptions of a PES. Making the assumption that individual-level (within-group) variables can be aggregated to form group-level (between-group) variables to draw inferences about group (e.g., school) qualities can be problematic and is often not a tenable assumption because the measurement structure may vary and between-group phenomenon may be unrelated to within-group phenomenon (Bliese, 2000; Longford & Muthén, 1992). While there is a long history of methodological research on cross-level invariance

[1] For example, U.S. Department of Education's Race to the Top (RTTT) competition, U.S. Department Teacher Incentive Fund grants (TIF), and state waivers for regulations in the No Child Left Behind Act all created incentives for states to adopt performance evaluation systems.

(e.g., Longford & Muthén, 1992; Selig et al., 2008), in education policy research and literature, cross-level invariance is often assumed rather than examined, even though researchers have demonstrated the implications of incorrectly assuming cross-level invariance (e.g., Kaplan, 2000; Schweig, 2014). As assuming cross-level invariance can lead to erroneous inferences and conclusions, to inform researchers about the group-level factor structure of teacher perceptions of PES, we examine the cross-level invariance. The results of the multilevel CFA provide some evidence that there is cross-level noninvariance, with fewer factors being identified at the school level, indicating that assumptions of cross-level invariance are problematic and that researchers should test multilevel invariance before grouping individual variables and making group inferences.

Furthermore, we examine the within-school agreement levels in teacher perceptions of PES. Based on business management literature (e.g., Bowen & Ostroff, 2004; Dickson, Resick, & Hanges, 2006; Lindell & Brandt, 2000), we argue that evaluation systems can help create a shared conception of good teaching. Agreement within schools about the fairness, credibility, and benefit of the evaluation system is a potential precursor to this development. A high level of agreement can be thought of as reflecting a strong evaluation climate within a school. To assess whether a strong evaluation climate had developed in schools within the district, we examine the agreement levels in teacher perceptions of the factors. While we find positive average perceptions for evaluator credibility and feedback quality, we find little evidence of an overall shared strong evaluation climate. Nonetheless, this approach illustrates the type of analysis that researchers, evaluators, or program administers could conduct to assess whether a strong evaluation climate is developing.

We first discuss a theory of action for PES, followed by a discussion of the relevant literature about teacher perceptions of PES, as well as cross-level measurement invariance and the implications of incorrectly assuming cross-level invariance. Furthermore, we review some business management literature that demonstrates how a strong evaluation climate may develop and be assessed. The research questions are presented, followed by an overview of the characteristics of the sample and the data and our methodology. The results, organized around the questions, are presented, and we conclude with a discussion of the results and the implications for future research.

Performance Evaluation System Theory of Action

For PES reforms to work, PES needs to be treated as a coherent system. Darling-Hammond (2013) argues that policymakers and practitioners need to think of teacher evaluation as a system with five elements (i.e., common standards, performance assessments guiding state functions, local evaluation systems aligned to same standards, aligned professional learning opportunities, and support structures) and that these elements all need to be in place for an evaluation system to be productive. We argue that in addition to establishing these program elements, teacher perceptions of the quality of these elements are just as critical for making PES work. For example, in addition to having a common set (state or local) of standards, teachers must generally agree that the standards are illustrative of "good" teaching practices. While local evaluation systems need to be aligned with the same standards and need to incorporate multiple valid and reliable measures of performance, in order for teachers to make changes to their actual practice, it is critical that teachers perceive the multiple measures as fair, valid, and reliable. In addition to having trained evaluators, teachers should perceive that their evaluators are knowledgeable, credible, and fair. The extent that teachers incorporate feedback from evaluators, pursue recommended professional development, and ultimately make changes in teaching practices is likely to some extent dependent on their perceptions of the fairness and validity of the multiple measures and their evaluator's credibility. Figure 1 illustrates one potential theory of action that links PESs to improved instruction and student learning.

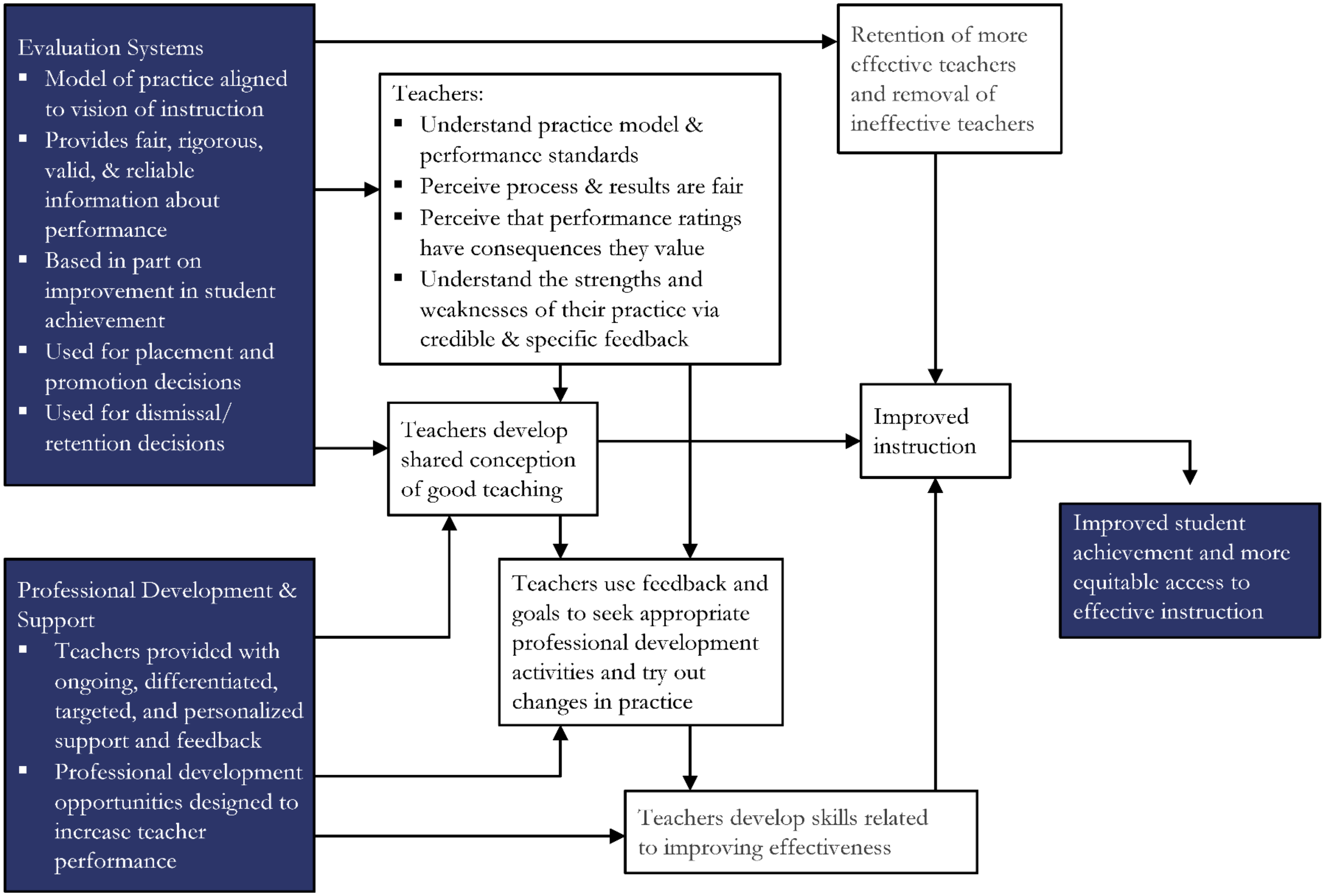
Evaluation Systems
▪ Model of practice aligned to vision of instruction
▪ Provides fair, rigorous, valid, & reliable information about performance
▪ Based in part on improvement in student achievement
▪ Used for placement and promotion decisions
▪ Used for dismissal/ retention decisions
Teachers:
▪ Understand practice model & performance standards
▪ Perceive process & results are fair
▪ Perceive that performance ratings have consequences they value
▪ Understand the strengths and weaknesses of their practice via credible & specific feedback
Retention of more effective teachers and removal of ineffective teachers
Teachers develop shared conception of good teaching
Improved instruction
Improved student achievement and more equitable access to effective instruction
Professional Development & Support
▪ Teachers provided with ongoing, differentiated, targeted, and personalized support and feedback
▪ Professional development opportunities designed to increase teacher performance
Teachers use feedback and goals to seek appropriate professional development activities and try out changes in practice
Teachers develop skills related to improving effectiveness

Figure 1. Performance Evaluation System Theory of Action

Relevant Literature

Recognizing the importance of teacher reactions and the linkages in the theory of action, many formative and summative evaluations of new teacher PESs include surveys of teacher perceptions (e.g., Firestone, Nordin, Shcherbakov, Kirova, & Blitz, 2014; Henry & Guthrie, 2015; Sporte & Jiang, 2016; Tennessee Department of Education, 2016). Several studies have examined the relationships between teachers' perceptions of evaluation and indicate that perceptions of understanding of the measures, fairness of the measures, evaluators' creditability, quality of the feedback, impact on professional development, and impact on instruction are all interrelated. To study educators' reactions to Chicago's Recognizing Educators Advancing Chicago Students (REACH) evaluation system, Jiang, Sporte, and Luppescu (2015) used survey and interview data to examine the relationship between teachers' perceptions of REACH, teacher characteristics, and perceptions of leadership and professional community. Jian et al. (2015) established scales representing the constructs of *Clarity*, *Practicality* (evaluation and feedback, PD and practice, and student growth), and *Cost*. The clarity measure was based on the extent that teachers understood components of the evaluation system, including professional practice rubric and student growth measures. The first practicality measure—*Evaluation and Feedback*—was based on teachers' perceptions of their evaluator's creditably (e.g., accurately assessing instruction, fair and unbiased) and on teachers' perceptions of the quality of the feedback they received (e.g., identified specific areas that could be improved, guidance on making improvements to instruction). The second practicality measure—*PD and Practice*—was based on teachers' perceptions of survey items asking about the extent that teachers used results to reflect on practice and to guide. The third practicality measure—*Student Growth*—was based on the extent that teachers felt the measures of growth were fair, would inform PD, and would lead to changes in instruction. The last construct—*Cost*—was based on teachers' perceptions that the evaluation increased levels of stress, relied too heavily on growth, and was more effort than it was worth. The authors found that *Clarity* and all of the *Practicality* constructs (*Student Growth*, *PD and Practice*, and *Evaluation and Feedback*) were moderately, positively related to teachers' perceptions of *Leadership* and *Professional Community*, with *Evaluation and Feedback* having the strongest relationship with *Leadership* and *Professional Community*. This study provides one example of how to begin to conceptualize and measure teachers' perceptions of a PES.

Cherasaro et al. (2016) used data from Regional Educational Laboratory Central's Examining Evaluator Feedback Survey to analyze teachers' perceptions of feedback provided by teacher evaluation systems. Based on Ilgen, Fisher, and Taylor's (1979) work on effective feedback and a review of research on performance feedback (e.g., Coggshall et al., 2012; Desimone, Porter, Garet, Yoon, & Birman, 2002; Kinicki, Prussia, Wu, & McKee-Ryan, 2004; Monk & King, 1994), the authors developed a theoretical model for performance feedback incorporating four interrelated components—*Usefulness of the Feedback*, *Accuracy of the Feedback*, *Evaluator Creditability*, and *Access to Resources*—that lead to changes in practice (referred to as response to feedback). Using structural equation modeling (SEM) analysis, the authors examined how characteristics of feedback and response to feedback are interrelated and found that teachers' responses to feedback is influenced by how useful they perceived it to be, which was influenced by their perceptions of the evaluators' credibility. They also found that teachers' perceptions of evaluator credibility is strongly related to their perceptions of accuracy of the feedback. This study offers another model for examining the relationship between teachers' perceptions of feedback and response changes to instructional practice, and it demonstrates the interrelations among teachers' perceptions of accuracy, evaluator credibility, usefulness, access to resources, and response to feedback. The authors concluded by calling for additional research to further examine the relationships between teachers' responses to feedback and actual changes in teachers' performance.

This article first examines the dimensions of interconnections among teachers' perceptions related to performance evaluation using survey data from a study of the implementation of a new PES in a large urban district. It uses SEM to examine the impacts of perceptions of system characteristics on teachers' perceptions of system impacts and the benefit of the process. This analysis is useful for two reasons. First, it helps establish the validity of the scales we developed to summarize teacher responses and measure the constructs of interest. It does this by providing evidence that the items are related to the constructs they were intended to measure, and by showing that the scales have the types of interrelationships one might expect if they were measuring constructs that theory and prior research both suggest should be causally related. That is, the interrelationships conform to the expected nomological network (Cronbach & Meehl, 1955). Second, it replicates prior research on teachers' perceptions and adds confirmation that perceptions of PES characteristics (e.g., fairness of measures, evaluator credibility, and quality of feedback) influence the impact of the evaluation process on teaching and teachers' perceptions of its potential benefits to improve student learning.

In addition to examining interconnections among teachers' perceptions at the individual level, we examine the multilevel nature of the survey data at the school level to assess whether there is cross-level measurement invariance (also referred to as equivalence). Researchers have demonstrated that there is cross-level noninvariance (nonequivalence) in the measurement models for working condition surveys and classroom environment surveys (Kaplan, 2000; Schweig, 2014), which can influence inferences about policy recommendations. For example, in a multilevel CFA, Kaplan (2000) demonstrated that a questionnaire with three sections—teacher quality, negative school environment, and student misbehavior—fit reasonably well with a three-factor at the student level (first level), but fit best with only one factor at the school level. That is, variation at the student level within schools differentiated between three dimensions, but only one factor explained between-school variation. In his analysis of the North Carolina Working Conditions Survey (WCS), Schweig (2014) found that accounting for the multilevel factorial structure of the WCS resulted in a different dominant factor of teacher attrition being identified versus when the multilevel factorial structure was ignored. These studies demonstrate that researchers and policymakers who assume cross-level invariance risk making incorrect assumptions about the number of distinct dimensions they are working with, which likely leads to a different set of policy recommendations and interventions. In the case of modeling teachers' perceptions of evaluation systems, at the school level fewer factors may be clearly distinguished, thus requiring a broader conceptualization of the factors at the school level.

The third focus of the article is an examination of the agreement in teacher perceptions within schools. We assert that in addition to the potential effects of PESs on improving individual teachers' practice (e.g., by providing feedback and motivating its use) and the overall quality of the workforce (e.g., by providing the basis for removing lower performers and recognizing high performers), PESs can contribute to creating a shared conception or vision of good practice that becomes embedded in the school culture. Such a vision provides a goal or standard for teachers to aspire to, day in and day out. Based on the management literature (e.g., Bowen & Ostroff, 2004; Dickson, Resick, & Hanges, 2006; Lindell & Brandt, 2000), we further argue that PESs can help create this shared conception.

One way that PESs could do so is by making evaluation a strong situation (Mischel, 1973), one in which events are perceived in the same way by the affected actors and expectations are clear (Schneider et al., 2002), and whereby ambiguity of organizational norms and practices is low, leading to more uniform perceptions and expectations for behavior (Dickson et al., 2006). Having shared perceptions of practices, policies, and routines, such as performance evaluation and its uses, promotes shared beliefs regarding how organizations operate and the norms and expectations for behavior. A strong situation "…creates consensus and similarity of perceptions among followers" (Feinberg et al., 2005, p. 472), so we would expect that where the PES has permeated the school, teachers will agree in their perceptions of the system, and if these perceptions are favorable, a consensus or "collective mindset" about expectations for performance will evolve. As a precursor to the development of a shared conception of good teaching, we would expect that teachers in schools tend to have similar (and of course positive) perceptions of understanding, fairness, evaluator credibility, and acceptance of the performance standards. Teachers' PES-related experiences and perceptions might be thought of constituting the school's evaluation climate, analogous to the way organizational researchers have conceptualized and studied safety climate (Griffin & Curcuruto, 2016) or customer service climate (Bowen & Schneider, 2014). A strong, positive evaluation climate could be a precursor of a shared conception of good teaching.

Because of the potential importance of a strong evaluation climate, this article examines within-school agreement in teacher perceptions. We focus on the school level because much of teachers' experience of evaluation is determined by what happens there. In many cases, the evaluators are school administrators who are likely to have their own approaches to observation, their own interpretations of rubrics, and their own styles of delivering feedback. Moreover, as shown by Halverson, Kelley, and Kimball (2004) and Halverson and Clifford (2006), school administrators can use evaluation systems to further their own agendas for managing their schools. Thus, we might expect some substantial differences between schools in how teachers perceive evaluation systems. District-wide survey results could therefore mask some substantial differences in how systems are perceived or experienced across schools. Evaluation system administrators might want to examine within-school agreement in perceptions as well as differences between schools in the average level of perceptions. Ideally, at least some schools should see both relatively high (and positive) average levels of perceptions of fairness, evaluator credibility, and feedback quality, in addition to high levels of agreement among teachers.

The analyses below address the following research questions:

Research Question 1: What is the factorial structure of teachers' perceptions of this performance evaluation system? What are the structural relationships between the identified factors?

Research Question 2: What is the multilevel factorial structure of teachers' perceptions of the performance evaluation system?

Research Question 3: How much agreement is there in teachers' perceptions of the performance evaluation system within the district's schools? Are there characteristics of schools that are associated with greater or lesser degrees of agreement?

District, Sample, and Data

In this analysis, we use data from a survey of teacher perceptions of a new PES implemented in a large, midwestern U.S. school district. This district had been piloting the new PES in the two years prior to the survey, during which time the system's procedures and performance measures were thoroughly developed and tested. The new PES was implemented in response to a change in state laws governing educator evaluation that required inclusion of measures of student achievement growth be taken into account in making teachers' summative performance ratings. A substantial amount of effort was expended to communicate about the PES to teachers and evaluators, and many teachers within most schools had experienced aspects of the PES in the phase-in period. The PES measured teachers' practice based on school administrators' classroom observations, using a rubric derived from the Framework for Teaching (Danielson, 2007). It also included measures of teacher effectiveness: classroom-level value-added for teachers of tested subjects, and a student learning objective for all teachers, with another for teachers for whom value-added could not be estimated (e.g., K-3 teachers; most high school teachers; and art, music, and PE teachers). School-level value-added was also included at varying weights, depending on the teaching assignment. These multiple measures were combined using formulae that differed by the type of student achievement growth data available for the teacher.

Data on teachers' perceptions of characteristics of the PES evaluation system and its impacts were collected as part of an evaluation of the implementation using an internet-based survey. The survey was administered in the spring of the first year of full implementation. The survey was designed in cooperation with a research organization that worked with the district, and it was embedded within a larger statewide survey of teacher working conditions. Because of this embedding, the number of questions that could be asked about the evaluation process had to be limited, to reduce overall respondent burden. The response rate for the entire survey within the focal district was approximately 80%; response rates for various items were considerably lower, however, ranging from 51% to 21%. This response rate resulted in a data set with 12,292 educators across 515 schools. Seventy-three percent were tenured, 26% non-tenured (typically teachers in their first four years with the district), and 1% were in other categories such as long-term substitutes. Eighty percent worked at traditional neighborhood schools, 16% at magnet or selective enrollment schools, and 4% in a variety of other programs. Teachers in charter schools were not included because these schools did not have to use the same evaluation system. The respondents were grouped at the school level with the average number per school being 24 (rounded to the nearest whole number).

This analysis focuses on a set of 26 survey items (Table 1) that were designed to measure teachers' perceptions of various aspects of a PES: understanding of the measures of the evaluation system, fairness of the measures, credibility of the evaluators, quality of the feedback received, impact of the evaluation system on collegiality, impact of the evaluation system on professional development (PD), impact of the evaluation system on teaching practices, and the benefits of the evaluation system. The survey items were on a Likert scale ranging from 1 to 4.

Table 1
Univariate Higher-Order Descriptive Statistics of Survey Items

Indicator	Indicator Text	*n*	*M*	Variance	ICC	Skew	Kurtosis
	How would you rate your understanding of…						
Q1	observation rubric or framework used to rate teacher professional practice.	12,208	2.89	0.71	0.06	-0.28	-0.64
Q2	how different assessments are combined to create growth measure.	12,172	2.71	0.81	0.05	-0.12	-0.81
Q3	how ratings of professional practice and student growth are combined to determine a summative performance.	12,163	2.70	0.81	0.05	-0.12	-0.82
	To what extent do you agree or disagree:						
Q4	The observation rubric is a fair representation of good teaching.	10,720	2.83	0.66	0.07	-0.48	-0.11
Q5	These measures of student achievement growth are a fair assessment of my students' learning.	10,075	2.44	0.71	0.06	-0.12	-0.64
Q6	The measures of student achievement growth based on performance tasks are a fair assessment of my students' learning.	10,077	2.49	0.69	0.04	-0.19	-0.57
	To what extent do you agree or disagree: My evaluator…						
Q7	knows my strengths and weaknesses as a teacher.	10,826	3.43	0.67	0.10	-1.39	1.22
Q8	knows what is going on in my classroom.	10,837	3.34	0.73	0.12	-1.18	0.57
Q9	is fair and unbiased.	10,830	3.50	0.69	0.09	-1.67	1.94

Table 1 (Cont'd.)
Univariate Higher-Order Descriptive Statistics of Survey Items

Indicator	Indicator Text	*n*	*M*	Variance	ICC	Skew	Kurtosis
	To what extent do you agree or disagree:						
Q10	The feedback I received this year identified specific areas of my instruction that could be improved.	10,752	3.31	0.55	0.08	-1.00	0.91
Q11	I have used the feedback I have received so far to improve my instruction.	10,746	3.38	0.52	0.09	-1.13	1.25
Q12	The feedback I received this year included guidance or suggestions on how to make improvements to my instruction.	10,736	3.30	0.61	0.09	-1.03	0.70
Q13	The observation process encouraged me to reflect on my teaching practice.	10,738	3.20	0.48	0.06	-0.77	1.05
Q14	My observation ratings will guide my future professional development choices.	10,714	3.02	0.61	0.06	-0.55	0.03
Q15	My evaluation results will strongly influence my future professional development activities.	10,691	2.96	0.64	0.07	-0.44	-0.27
Q16	The information I get from these assessments will inform my professional development choices.	10,052	2.73	0.69	0.09	-0.45	-0.24
Q17	The information I get from the performance tasks will inform my professional development choices.	10,071	2.70	0.67	0.07	-0.41	-0.26

Table 1 (Cont'd.)
Univariate Higher-Order Descriptive Statistics of Survey Items

Indicator	Indicator Text	*n*	*M*	Variance	ICC	Skew	Kurtosis
Q18	I have made changes in my teaching as a result of the observation process.	10,745	3.15	0.50	0.06	-0.70	0.77
Q19	I have made changes in my teaching in order to improve my student scores on these assessments.	10,056	3.05	0.59	0.10	-0.79	0.70
Q20	I have made changes in my teaching in order to improve my student scores on type III assessments.	10,076	2.97	0.58	0.08	-0.72	0.61
Q21	The evaluation process has improved the quality of my conversations with my colleagues about instruction.	5,854	2.74	0.65	0.08	-0.29	-0.33
Q22	The evaluation process encourages teachers to collaborate.	5,863	2.74	0.68	0.08	-0.28	-0.42
Q23	The evaluation process has improved my communication with leadership at this school.	5,817	2.76	0.70	0.08	-0.35	-0.37
Q24	The evaluation process will lead to better instruction.	11,997	2.74	0.64	0.07	-0.46	-0.11
Q25	The evaluation process will lead to improved student learning.	11,969	2.70	0.66	0.07	-0.35	-0.29
Q26	The evaluation process takes more effort than the results are worth.	11,903	2.73	0.71	0.04	-0.02	-0.76

Note. n=12,292. Clustered at school level. Number of clusters is 515. Average cluster size is 23.868. ICC = intraclass correlation. ICC values of ≥0.10 are shaded gray.

Methods

We conducted several different analyses to address our research questions. For the first question, we conducted a confirmatory factor analysis (CFA) and structural equation modeling (SEM). For the second question, we conducted a multilevel SEM. For the third question, we analyzed the variation within- and between-schools in the level of the eight scales using primarily descriptive statistics.

Our CFA and SEM models incorporate several measures and constructs similar to Jiang et al.'s (2015) and expand on Cherasaro et al.'s (2016) by examining the factorial validity of additional constructs/factors and examining more of the structural relationships between the factors. We begin by confirming the factorial validity of the following factors using CFA: *Understand Measures* (factor 1), *Measure Fairne*ss (factor 2), *Evaluator Credibility* (factor 3), *Feedback Quality* (factor 4), *Impact on Professional Development* (factor 5), *Impact on Teaching* (factor 6), *Impact on Collegiality* (factor 7), and *Evaluation Beneficial* (factor 8). Then we assess the structural relationships between the factors using SEM.

After we assess the structural relationships between the factors, we examine the multilevel structure of the factors. The CFA and SEM models discussed above examine the factorial validity of constructs and their structural relationships to each other at the individual level and provide information about between individual variables. However, as discussed above, researchers have demonstrated that there is cross-level noninvariance (nonequivalence) in the measurement models for working condition surveys and classroom environment surveys (Kaplan, 2000; Schweig, 2014), which can influence inferences about policy recommendations. In the present case, while variation at the individual level may differentiate across eight dimensions, at the school level fewer factors may explain between-school variation. To examine the factorial structure of teachers' perceptions of PES between-school variables, we conducted a multilevel CFA following procedures detailed in Byrne (2012) for multilevel modeling.

In conducting a multilevel CFA, one of the first steps is to inspect the intra-class correlation (ICC) values to determine if the amount of variance at the between-group level warrants a multilevel factor analysis (Muthén, 1994; Reise, Ventura, Nuechterlein, & Kim, 2005). Muthén (1997) recommends that when group sizes exceed 15 and findings yield ICC values of 0.10 or larger, the multilevel structure of the data should be accounted for in the modeling, but more recently Selig et al. (2008) recommended that even ICC values of less than .10 should not be ignored. The ICC values range from .04 to .12, with two of the 26 ICC values being greater than .1, and several at .09 (Table 1).

The CFA, SEM, and multilevel CFA were ran using pairwise option in Mplus. Missing data was not imputed. CFAs and SEMs were estimated in Mplus version 7.4 (Muthén & Muthén, 2015). Maximum likelihood (ML) was used as the estimator. After the initial hypothesized models were estimated, the modification indices were reviewed to determine additional parameters in the model based on substantive meaning, whether the expected parameter change (EPC) statistics were substantial, and the extent that the existing model exhibited adequate fit statistics. The goodness-of-fit statistics for the CFA (model 1) and SEM (model 2) are in the well-fitting ranges (Table 2). The root mean square error of approximation (RMSEA) values are below .05, the Comparative Fit Index (CFI) and Tucker-Lewis Fit Index (TLI) values are above .95, and the standardized root mean square residual (SRMR) values are .05 or below, all of which indicate a well-fit model (Browne & Cudeck, 1993; Byrne, 2012; Hu & Bentler, 1999). The goodness-of-fit statistics for the multilevel CFA (model 3) are in the well-fitting range, except for the between-level SRMR statistic, which is in the poor-fit range. The SRMR represents the average discrepancy between the observed sample and hypothesized correlation matrices in a standardized form, which can be interpreted as meaning that model 3 explains the correlations to within an average error of .14 at the school level (between-level). The within- and between-level SRMR values for model 3 suggest that the model fits the data better at the individual (within) level than the school (between) level.

Table 2
Fit Indices for CFA and SEM Models of Educators' Perceptions of Evaluation System

	Model 1: CFA	Model 2: SEM	Model 3: Multilevel CFA
Number of free parameters	116	98	179
Chi-Square Test of Model Fit			
Value	5881.43***	7801.50***	6718.58***
DF	261	279	549
Chi-Square Test of Model Fit for the Baseline Model			
Value	224543.61***	224543.61***	203052.65***
DF	325	325	650
Loglikelihood			
H0 Value	-213570.11	-214530.15	-212361.13
H1 Value	-210629.39	-210629.39	-209167.06
Information Criteria			
Akaike (AIC)	427372.22	429256.29	425080.25
Bayesian (BIC)	428232.56	429983.13	426407.84
Sample-Size Adjusted BIC	427863.92	429671.70	425839.00
Root mean square error of approximation (RMSEA)[1]			
Estimate	.042	.047	.030
90% C.I.	.041–.043	.046–.048	
Probability RMSEA <= .05	1	1	
Comparative Fit Index (CFI)[2]	.975	.966	.970
Tucker-Lewis Fit Index (TLI)[3]	.969	.961	.964
Standardized root mean square residual (SRMR)[4]	.046	.05	
Within			.046
Between			.137

Note. n=12,292. CFA, SEM, and multilevel CFA models each have 10 crossloadings.[1] RMSEA values < .05 indicate good fit (Browne & Cudeck, 1993).[2] CFI values of 0.90 to 0.95 are indicative of acceptable fit (Bentler, 1990) and values > .95 are indicative of a well-fit model (Hu & Bentler, 1999).[3] TLI values of .90 to .95 are indicative of acceptable fit (Bentler, 1990) and values > .95 are indicative of a well-fit model (Hu & Bentler, 1999).[4] SRMR values < .05 are indicative of a well-fit model (Byrne, 2012).
*p < .05. **p < .01. ***p < .001.

Furthermore, to assess degrees of school level agreement and makes inferences about the strength of the PES climate at the school level, we examined the descriptive statistics and variance of the teacher perception scales. Organizational researchers (e.g., Aksoy & Bayazit, 2014; Lindell & Brandt, 2000; Van Vianen et al., 2014) have argued that climate strength can be characterized in terms of both the level of perceptions (e.g., the average of group members' perceptions) and the agreement among group members. The level of perception has been viewed as an indicator of the quality of the climate, as well as the agreement as an indicator of the consensus or the degree to which the perception is shared. A strong climate exists when the level of positive perceptions is high and the degree of agreement is high as well.

Results

Dimensions of and Interrelationships Among Teachers' Perceptions of a Performance Evaluation System

The CFA model identified eight factors with all indicators significantly loading onto the factors and estimates ranging from -.32 to .96 (Table 3). The results also indicate that that all of the factors are moderately to strongly related to each other, with standardized factor loadings ranging from .40 (relationship between *Evaluator Credibility* [factor 3] and *Understand Measures* [factor 1]) to .97 (relationship between *Impact on Teaching* [factor 6] and *Impact on PD* [factor 5] (Table 4). After testing the theoretical constructs using the CFA, we tested the validity of a causal structure using SEM. Figure 2 depicts the SEM model (model 2).

Model 2 estimates structural pathways for: (1) Understand Measures (F1) to Measure Fairness (F2), (2) Measure Fairness (F2) to Evaluator Credibility (F3), (3) Measure Fairness (F2) and Evaluator Creditability (F3) to Feedback Quality (F4), (4) Measure Fairness (F2) to Impact on PD (F5) and Impact on Collegiality (F7), (5) Feedback Quality (F4) and Impact on PD (F5) to Impact on Teaching (F6), (6) Impact on Teaching (F6) and Impact on Collegiality (F7) to Evaluation Benefits (F9). This model indicates that teachers' understanding of the evaluation measures (F1) has a direct effect on teachers' perceptions of the measures' fairness (F2), which have a direct effect on the perceived credibility of evaluators (F3). Teachers' perceptions of the measures' fairness (F2) and evaluators' credibility (F3) have a direct effect on teachers' perceptions of the quality of the feedback received from evaluation process (F4). Teachers' perceptions of the measures' fairness (F2) also have a direct effect on future choices and activities for PD (F5) and collegiality (F7). Teachers' perceptions of the quality of the feedback received as part of the evaluation process (F4) and influence on PD (F5) have a direct effect on changes in teaching practices (F6). Changes in teachers' practices (F6) and collegiality (F7) have a direct effect on overall perceptions of the benefits of the PES (F8).

The structural model parameter standardized estimates range from .18 (F6 on F4) to .86 (F7 on F2). Interestingly, but not unsurprisingly, *Measure Fairness* (F2) is directly significantly related to multiple other factors, including *Evaluator Creditability* (F3) (STDYX standardized coefficient = .63, SE = .01, p < .001), *Feedback Quality* (F4) (STDYX standardized coefficient =.34, SE = .01, p < .001), and *Impact on Collegiality* (F7) (STDYX standardized coefficient =.86, SE = .01, p < .001). Regarding *Impact on Teaching* (F6), the standardized path coefficient value is larger for *Impact on PD* (F5) than for *Feedback Quality* (F4) (STDYX standardized coefficient = .85 versus .18), indicating PD choices and activities are more strongly associated with changes in teacher practices than direct feedback as part of the PES. Also, *Impact on Collegiality* (F7) has a larger standardized (STDYX) regression coefficient (.62) on *Evaluation Benefits* (F8) than *Impact on Teaching* (F6) (standardized (STDYX) regression coefficient = .22), indicating that changes in collaboration and communication

Table 3
Standardized (STDYX) Factor Loadings Estimates for CFA

Factor/indicator	Estimate	S.E.	Est./S.E.
Understand measures (F1) by			
Q1	.79***	.004	205.80
Q2	.91***	.002	378.00
Q3	.94***	.002	437.96
Measure fairness (F2) by			
Q4	.78***	.006	138.84
Q5	.57***	.008	75.48
Q6	.56***	.008	74.54
Evaluator credibility (F3) by			
Q7	.90***	.003	342.33
Q8	.89***	.003	312.31
Q9	.82***	.004	215.49
Feedback quality (F4) by			
Q10	.90***	.002	384.26
Q11	.92***	.002	459.77
Q12	.92***	.002	462.98
Impact on professional development (F5) by			
Q13	.81***	.004	210.14
Q14	.87***	.003	258.90
Q15	.85***	.004	241.68
Q16	.57***	.007	79.87
Q17	.59***	.007	85.64
Impact on teaching (F6) by			
Q18	.86***	.005	157.46
Q19	.46***	.008	55.34
Q20	.49***	.008	60.49
Impact on collegiality (F7) by			
Q21	.90***	.003	265.70
Q22	.87***	.004	221.29
Q23	.85***	.004	195.27
Evaluation beneficial (F8) by			
Q24	.96***	.002	573.84
Q25	.96***	.002	555.74
Q26	-.32***	.008	-37.65

Note. *n*=12,292. Estimates are STDYX standardized, which is based on background and outcome variables. "By" is short for "measured by" and is used to indicate the regression estimate between the underlying latent factors (F1-F8) and the observed indictor variables. Model includes 10 crossloadings. Complete indicator text is available in Table 1.
*p < .05. **p < .01. ***p < .001.

Table 4

Standardized (STDYX) Factor Covariance Estimates for CFA

Factor	Estimate	S.E.	Est./S.E.
Measure fairness (F2) with			
Understand measures (F1)	.59***	.009	68.65
Evaluator credibility (F3) with			
Understand measures (F1)	.40***	.009	45.23
Measure fairness (F2)	.60***	.009	67.14
Feedback quality (F4) with			
Understand measures (F1)	.43***	.008	50.08
Measure fairness (F2)	.61***	.009	70.57
Evaluator credibility (F3)	.77***	.005	158.33
Impact on professional development (F5) with			
Understand measures (F1)	.51***	.008	62.51
Measure fairness (F2)	.88***	.006	142.03
Evaluator credibility (F3)	.59***	.008	77.68
Feedback quality (F4)	.69***	.006	111.82
Impact on teaching (F6) with			
Understand measures (F1)	.47***	.009	49.94
Measure fairness (F2)	.77***	.009	86.77
Evaluator credibility (F3)	.57***	.009	65.03
Feedback quality (F4)	.74***	.007	106.29
Impact on professional development (F5)	.97***	.006	173.40
Impact on collegiality (F7) with			
Understand measures (F1)	.49***	.010	51.56
Measure fairness (F2)	.78***	.009	87.65
Evaluator credibility (F3)	.53***	.010	53.17
Feedback quality (F4)	.57***	.009	62.08
Impact on professional development (F5)	.78***	.007	115.07
Impact on teaching (F6)	.73***	.009	76.79
Evaluation beneficial (F8) with			
Understand measures (F1)	.44***	.008	55.13
Measure fairness (F2)	.84***	.006	136.65
Evaluator credibility (F3)	.46***	.008	54.29
Feedback quality (F4)	.48***	.008	60.89
Impact on professional development (F5)	.72***	.006	125.86
Impact on teaching (F6)	.66***	.008	84.25
Impact on collegiality (F7)	.76***	.006	123.76

Note. n=12,292. Estimates are STDYX standardized, which is based on background and outcome variables. "With" is short for "correlated with" and is used to indicate covariance relations between latent variables in the measurement model.

$^{*}p < .05$. $^{**}p < .01$. $^{***}p < .001$.

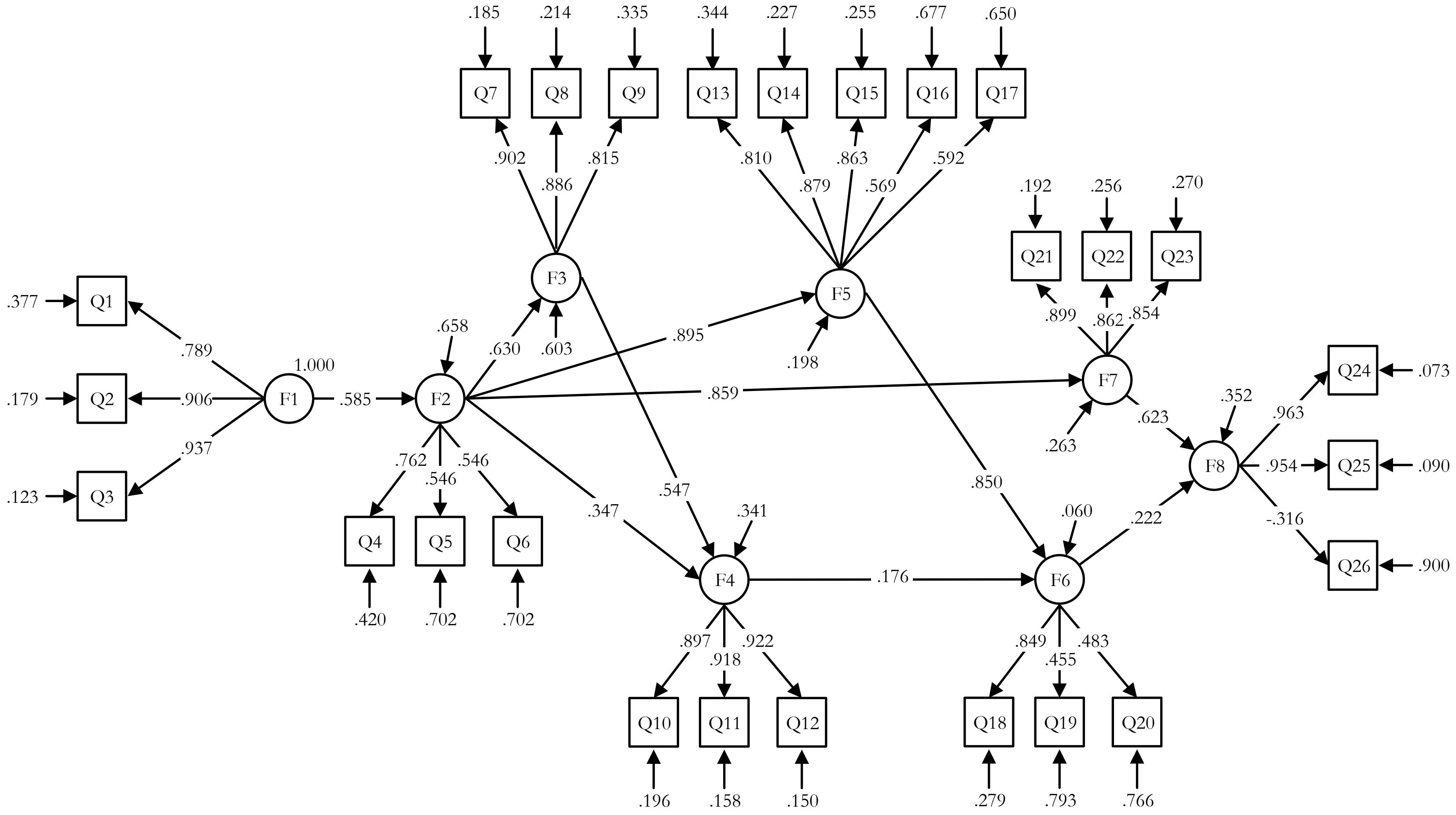

Figure 2. Structural Equation Model of Interrelationships between Teacher Perceptions of a Performance Evaluation System

Note. n=12,292. Estimates presented are STDYX standardized, which is based on background and outcome variables. The single-headed arrows leading from each factor to the related indicators indicate the regression estimates of each item onto the underlying factor (i.e., factor loadings). The single-headed arrows from one factor to another indicate the regression estimate of one factor onto another. The single-headed arrows pointing to each of the observed variables indicate measurement error (i.e. residuals). The short single-headed arrows pointing to each of the factors indicate the residual variances.
All parameter estimates are significant at $p < .001$.

with colleagues are more strongly associated with perceived benefits of the evaluation system (including better instruction and improved student learning) than changes in teachers' instructional practice.

Initial School Level Agreement: Evaluation Climate Strength Assessment

We begin examining climate strength by analyzing the variation within and between schools in the level of eight of the scales that were used in the SEM analysis above. We created factor-based scales by averaging across the items used in the SEM to preserve the original metric of the items. Table 5 shows the variance decomposition between the school and individual levels and the average reliability of a school average.

Table 5
Teacher Perception Scale Variance Between and Within Schools

Scale	Proportion of variance between schools	Proportion of variance within schools	Reliability of school average
1. Understand Measures	.06*	.94	.60
2. Measure Fairness	.06*	.94	.60
3. Evaluator Credibility	.11*	.89	.72
4. Feedback Quality	.10*	.90	.69
5. Impact on Professional Development	.08*	.92	.68
6. Impact on Teaching	.08*	.92	.69
7. Impact on Collegiality	.08*	.92	.52
8. Evaluation Beneficial	.08*	.92	.68

Note. *p < .05. **p < .01. ***p < .001.

The proportion of variance between schools is also the intra-class correlation (ICC), which is often interpreted as a measure of within-group agreement as well as an estimate of the proportion of variance in scale scores that lies at the group level. The table shows that the vast majority of variation in scale values is at the individual (within-school) level, though the proportion of variance between schools is not negligible. These schools did differ in the average favorability of teacher perceptions about these aspects of the evaluation process, and the reliability of the school averages is high enough to consider between-school analyses. All ICCs are significant at the .05 level or beyond. However, when interpreted as a measure of within-school agreement, they suggest relatively low within-school agreement.

One potential problem with the ICC as a measure of within-school agreement is that its value depends on the magnitude of the between-group variance as well the similarity of responses within groups (Lindell & Brandt, 2000). School-level variance could be low if districts implemented evaluation systems uniformly across schools and schools were implementing in pretty much the same way. If all schools were doing a "good job," then we would not expect to see a lot of inter-school variation. In this district, a lot of resources were expended on implementation, including training, communication, and infrastructure for teachers and evaluators to support the process. Therefore, it might be the case that teacher perceptions do not differ substantively across schools. In this case, the relatively low proportion of variance at the school level suggests that the ICCs might underestimate within-school agreement. If teachers' perceptions are largely influenced by idiosyncratic factors not widely shared within the school, we would expect to find that most variation is within rather than between schools. However, some school characteristics might also be associated with the level of agreement. We next turn to a different measure of within-school agreement and explore some potential influences on within-school agreement.

Both to separate within-group agreement from the extent of between-group variation and to enable examination of agreement within individual groups, Lindell and Brandt (2000) recommended calculating an index of agreement such as the mean absolute deviation (the absolute value of the difference between the individual value and the group mean) that assesses agreement within groups independently of the degree of between-group variation. We chose to calculate the mean absolute deviation because its interpretation is more transparent than the standard deviation or other, more specialized indices. Table 6 shows the average mean absolute deviations for each scale at the school level and the 10th and 90th percentiles of these averages, and it provides the average individual-level mean absolute deviation for each scale.

Table 6
Average Individual-Level and Within-School Mean Absolute Deviations for Scales

Scale	Average Within-School Mean Absolute Deviation	10th and 90th Percentile Within-School Mean Absolute Deviation	Individual-Level Average Mean Absolute Deviation
1. Understand Measures	.62	.47 - .75	.63
2. Measure Fairness	.52	.38 - .65	.53
3. Evaluator Credibility	.54	.29 - .79	.55
4. Feedback Quality	.52	.27 - .69	.52
5. Impact on PD	.46	.33 - .59	.48
6. Impact on Teaching	.43	.30 - .55	.58
7. Impact on Collegiality	.52	.29 - .73	.54
8. Evaluation Beneficial	.48	.33 - 61	.49

Note. For schools with five or more respondents, *n*=496.

The average within-school mean absolute deviations are relatively large and almost as large as the mean absolute deviations calculated based on individual responses. This substantiates the impression received from reviewing the ICCs that school effects are on average small. Yet there were substantial differences among schools. Appendix B, Figure 1 shows the distributions of school mean absolute deviations for each scale. There were clearly some schools with relatively high levels of agreement (.1 to .3 mean absolute deviations) and some with much lower levels (MAD of .7 or above). Interestingly, evaluator credibility and feedback quality, which had the highest ICC values of the eight scales, did not have the lowest average absolute mean deviation. We expected these two to have lower average mean absolute deviations, because they refer to aspects of the evaluation situation that school administrators can strongly influence. Understanding of the performance measures and perceptions of their fairness had higher average mean absolute deviations, as one would expect given that measures were chosen and communicated at the district level, and because the student achievement growth measures varied across teacher assignments.

One might hypothesize that schools with strong or weak consensus on perceptions of one characteristic of evaluation might also have strong or weak consensus on others. If performance evaluation is to present a strong situation, teachers within a school should tend to agree in their perceptions of multiple important aspects of the system. Somewhat surprisingly, in this district the mean absolute deviations were not strongly correlated across most scales. Table 7 shows the correlation between school level mean absolute deviations. The strongest correlations (*r* >= .5) are shaded.

Table 7
Correlations Between School Average Mean Absolute Deviations

Scale	1	2	3	4	5	6	7	8
1. Understand Measures	1							
2. Measure Fairness	.33	1						
3. Evaluator Credibility	.15	.13	1					
4. Feedback Quality	.19	.14	.60	1				
5. Impact on PD	.29	.55	.16	.24	1			
6. Impact on Teaching	.20	.34	.06	.21	.56	1		
7. Impact on Collegiality	.30	.32	.15	.15	.38	.29	1	
8. Evaluation Beneficial	.20	.44	.11	.15	.38	.29	.42	1

Note. For schools with five or more respondents, *n*=496.

As might be expected given that the underlying items will often pertain to the same person, the deviations for evaluator credibility are correlated with those for feedback quality, indicating that in schools where agreement was lower on evaluator credibility, it was also lower on feedback quality. In schools with less agreement about the impact of the evaluation process on professional development, there tended to be less agreement on the impact on teaching. Schools with less agreement about the fairness of the measures also had less agreement about the evaluation system's impact on professional development.

We next assessed whether there were school characteristics that were associated with agreement. One complicating factor is that the degree of agreement is inherently correlated with average level of the perceptions (Cole et al., 2011; Lindell & Brandt, 2000). To attain a high average level, all or most group members have to give high-valued responses. Similarly, when group averages are very low, all or most members provide low-valued responses. Thus, group average levels and intra-group agreement do not provide independent information at the high and low levels of measures of evaluation system perceptions. This dilemma has led researchers trying to connect perceptions of climate to measures of unit performance to include both the main effects of level and agreement, plus an interaction term. Studies relating climate or other group-level constructs measured by individual perceptions to outcomes have been mixed as to whether adding an agreement measure and the interaction lead to better prediction of unit performance (Feinberg et al., 2005; Gonzoles-Roma et al., 2002; Lindell & Brandt, 2000; Schneider et al., 2002; Van Vianen et al., 2011). It appears that for some, but not all, measures of different types of climate, adding a consensus measure explains at best a modest amount of additional variation in various group performance measures (e.g., innovation, organizational commitment, customer perceptions of service quality).[2]

Although we do not have data on outcomes that can be related to perceptions of evaluation systems and the consensus of those perceptions, we did examine the degree to which the level of the scales was related to the within-school agreement. As expected, schools with higher and lower scale averages did have lower mean absolute deviations, at least for some of the scales. Figure 3 shows the relationship between the mean average deviations and the school-level scale means for four of the scales, which represent the four patterns we found. In most cases, as represented by the *Understand Measures* scale, there was a small to moderate tendency for the mean absolute variations to be lower at high and low values (school averages). In two cases—the *Evaluator Credibiity* scale (shown below) and the similar-appearing *Feeback Quality* scale (not shown below)—there was a strong tendency for consensus to increase as school average scores go up, due in part to the relatively high averages for these scales. In two cases, exemplified by the *Impact on PD* and *Evaluation Beneficial* scales, consensus increases slightly as school averages go up.

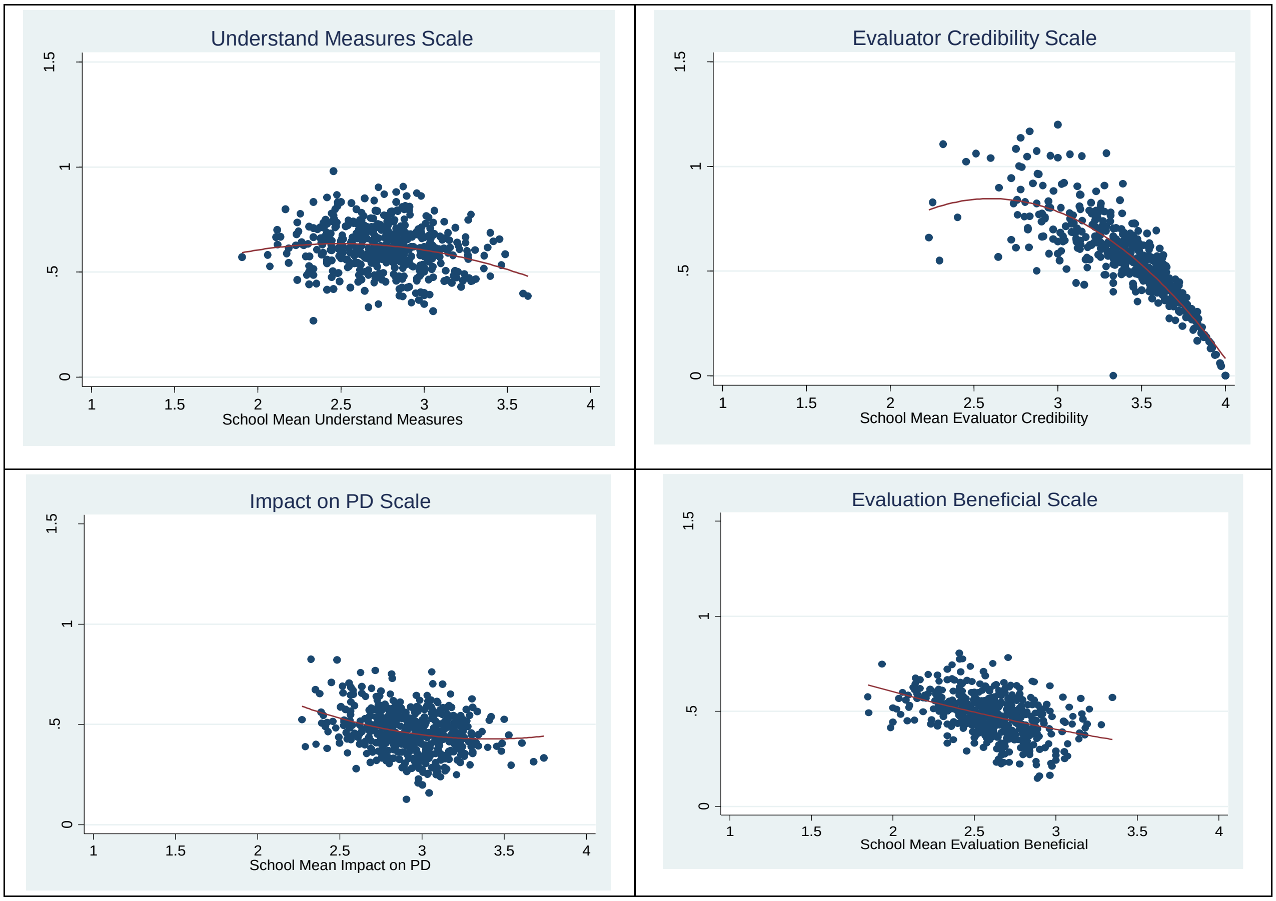

Figure 3. The Relationship of School Level Mean Absolute Deviations and Average Scale Levels
Note. For schools with five or more respondents, n=496.

[2] Interestingly, Schweig (2016) found that consensus in student ratings of teachers was related to teacher value-added, in addition to the average rating given by the students.

It appears that schools with higher average values for the scales tend to have less disagreement on most scales, as shown by the downward slope of the fitted lines in Figure 3. However, the strength of this tendency varies across scales, and there is still substantal variation in agreement around the average of the scales. The relationship between evaluator credibility and feedback quality means and the mean absolute deviations is especially strong, perhaps because these two scales focus on an evaluator many teachers in a school are likely to share (e.g., the principal or an assistant principal).

Correlates of Agreement

Evaluation system administrators are likely to be interested in whether certain types of schools have greater of lesser teacher consensus. Several school characteristics could be related to consensus in teacher perceptions of evaluation. Both larger schools and high schools (typically with more specialized and departmentalized teachers) could have less agreement because of more limited interaction among teachers and differing disciplinary perspectives. Teachers in schools in NCLB corrective action or restructuring status might feel greater pressure to perform, or they might have principals who are polarizing the teaching staff by introducing reforms or trying to weed out lower performers. This might also be the case in schools with relatively low student proficiency. As a consequence, these schools could have more disagreement about the evaluation process. We also examined whether schools with a higher proportion of tenured teachers might have higher average mean absolute deviations, since tenured teachers had less reason to pay attention to the evaluation process, and non-tenured teachers were more likely to have had exposure to the new system in prior years, as well as being less used to the old evaluation system. Table 8 shows the partial correlations between each of these characteristics and schools' average mean absolute deviation for each of the eight scales. Partial correlations were calculated to remove the effects of level, and because in most cases there was a curvilinear relationship between mean absolute deviation and level, both level and the square of level were partialed out.

As the table shows, only school size has a consistently positive relationship with average disagreement, and the relationship is relatively weak. The relationship with size seems reasonable, because it is likely easier to achieve consistent implementation in smaller schools, and in smaller schools one administrator is more likely to handle the evaluation. The proportion of tenured teachers is related to disagreement for seven of the eight scales, but the relationship is small and only significant for one scale.

Table 8
Relationship of School Average Mean Absolute Deviations to School Demographic Characteristics

Scale	Partial correlation of school average MAD with:					
	Log school size	High school	% low income	% of tenured teachers	% proficient, prior year	School in corrective action or restructuring
1. Understand Measures	.16*	.07	-.01	.09*	-.02	.04
2. Measure Fairness	.21*	.12*	.05	.06	-.05	.04
3. Evaluator Credibility	-.01	-.07	.08	.05	-.02	.11*
4. Feedback Quality	.11*	-.08	-.04	-.01	.08	-.05
5. Impact on PD	.20*	.08	.00	.06	.07	-.03
6. Impact on Teaching	.11*	.08	-.04	.03	.01	-.03
7. Impact on Collegiality	.12*	.10*	.01	.06	-.02	-.03
8. Evaluation Beneficial	.14*	-.02	-.02	.08	.03	-.03

Note. For schools with five or more survey respondents.
*p < .05. **p < .01. ***p < .001.

We also examined the relationship between the school average mean absolute deviations and four scales representing teachers' perceptions of school leadership.[3] These measures were taken from the broader survey in which the evaluation system perceptions were also collected, so there is likely to be some common method and common situation effects that could bias estimated correlations upward. Table 9 shows the partial correlations.

[3] See Appendix C for examples of the items making up these scales.

Table 9

Relationship of School Average Mean Absolute Deviations to School Climate Measures

Scale	Partial correlation of school average MAD with:			
	Teacher-principal trust	Principal instructional leadership	Program coherence	Teacher influence
1. Understand Measures	.05	.06	.00	.05
2. Measure Fairness	.04	.07	.03	.04
3. Evaluator Credibility	.09*	.15*	-.09	.07
4. Feedback Quality	.07	.08	.05	.07
5. Impact on PD	.06	.06	.02	.02
6. Impact on Teaching	.09	.01	.07	.02
7. Impact on Collegiality	-.01	.05	.01	-.01
8. Evaluation Beneficial	.11*	.14*	.12*	.09*

Note. For schools with five or more survey respondents.
$^*p < .05$. $^{**}p < .01$. $^{***}p < .001$.

It was surprising that the significant positive correlations seem to be in the wrong direction. One would expect that stronger principal leadership and teacher-principal trust would increase within-school consensus. It may be that these general measures of leadership in schools do not represent aspects of principal behavior related to evaluation.

Cross-Level Measurement Invariance

Evaluation system administrators are likely to be interested in which schools appear to be developing a strong evaluation climate. A strong climate requires both high average values on the scales and high consensus (a small absolute mean deviation). One difficulty with trying to identify these schools is that it might not be appropriate to compare schools' average factor scores developed based on a model that was based on both within-school and between-school variation. Since much of the variation was at the individual teacher (within-school) level, we recognized that factors based on the SEM analysis discussed above might not show up when making between school comparisons. Factors that are based primarily on individual level variance could be different than those that would be revealed by an analysis of between-school variance. Thus, we next assessed the cross-level measurement invariance using multilevel confirmatory factor analysis.

Cross-level invariance was first assessed using a multilevel CFA that replicated the factorial structure at the individual and group level (i.e., an eight-factor model). While the model estimation terminated normally, the residual covariance matrix was not positive definite, indicating the model needed to be modified. We then estimated a model with eight individual level factors corresponding to those used in the SEM (F1-F8) and two factors at the school (between) level (BF1), which we termed prerequisites, and BF2, which we called Impacts). As discussed above, the model fit indices are within the well-fitting range except for the between level SRMR, which, at a value of .14, indicates poor fit. This SRMR value indicates that the model does not explain the correlations well for the between model. Consistent with multilevel findings, the factor loadings are larger at the between-group level than at the individual level (see Table 10).

Table 10
Multilevel CFA Standardized (STDYX) Estimates for Within and Between Level

Factor/variable	Estimate	S.E.	Est./S.E.
Within level			
Understand measures (F1) by			
Q1	.78***	.007	120.15
Q2	.90***	.004	244.41
Q3	.93***	.003	274.78
Measure fairness (F2) by			
Q4	.78***	.007	107.07
Q5	.55***	.012	47.99
Q6	.55***	.011	50.36
Evaluator credibility (F3) by			
Q7	.90***	.004	245.10
Q8	.88***	.005	177.13
Q9	.81***	.008	103.32
Feedback quality (F4) by			
Q10	.89***	.005	180.39
Q11	.91***	.004	253.78
Q12	.91***	.004	236.69
Impact on Professional Development (F5) by			
Q13	.81***	.006	134.03
Q14	.86***	.005	166.08
Q15	.84***	.006	143.11
Q16	.54***	.012	45.28
Q17	.57***	.011	49.61
Impact on teaching (F6) by			
Q18	.85***	.008	105.62
Q19	.44***	.013	34.90
Q20	.47***	.012	37.90
Impact on collegiality (F7) by			
Q21	.90***	.005	168.65
Q22	.86***	.006	139.08
Q23	.84***	.007	128.95

Table 10 (Cont'd.)

Multilevel CFA Standardized (STDYX) Estimates for Within and Between Level

Factor/variable	Estimate	S.E.	Est./S.E.
Evaluation beneficial (F8) by			
Q24	.96***	.002	454.03
Q25	.95***	.003	372.21
Q26	-.29***	.013	-21.79
Measure fairness (F2) with			
Understand measures (F1)	.57***	.010	56.13
Evaluator credibility (F3) with			
Understand measures (F1)	.36***	.011	34.01
Measure fairness (F2)	.58***	.011	51.10
Feedback quality (F4) with			
Understand measures (F1)	.39***	.011	35.34
Measure fairness (F2)	.59***	.010	58.84
Evaluator credibility (F3)	.74***	.008	94.02
Impact on professional development (F5) with			
Understand measures (F1)	.48***	.010	49.28
Measure fairness (F2)	.87***	.008	110.05
Evaluator credibility (F3)	.57***	.011	53.07
Feedback quality (F4)	.68***	.009	75.16
Impact on teaching (F6) with			
Understand measures (F1)	.44***	.011	39.07
Measure fairness (F2)	.75***	.011	69.41
Evaluator credibility (F3)	.55***	.013	42.92
Feedback quality (F4)	.73***	.011	69.19
Impact on professional development (F5)	.97***	.008	116.08
Impact on collegiality (F7) with			
Understand measures (F1)	.46***	.012	40.30
Measure fairness (F2)	.77***	.011	69.92
Evaluator credibility (F3)	.50***	.013	38.23
Feedback quality (F4)	.54***	.011	48.14
Impact on professional development (F5)	.76***	.009	81.67
Impact on teaching (F6)	.70***	.013	55.08

Table 10 (Cont'd.)
Multilevel CFA Standardized (STDYX) Estimates for Within and Between Level

Factor/variable	Estimate	S.E.	Est./S.E.
Evaluation beneficial (F8) with			
Understand measures (F1)	.41***	.010	42.14
Measure fairness (F2)	.83***	.008	102.35
Evaluator credibility (F3)	.43***	.012	36.92
Feedback quality (F4)	.45***	.010	46.18
Impact on professional development (F5)	.69***	.008	83.49
Impact on teaching (F6)	.63***	.011	58.29
Impact on collegiality (F7)	.74***	.010	77.17
Between level			
Prerequisites (BF1) by			
Q1	.89***	.029	31.23
Q2	.99***	.009	111.14
Q3	1.0***	.011	94.00
Q4	.81***	.038	21.64
Q5	.70***	.057	12.21
Q6	.70***	.059	11.96
Q7	.97***	.010	96.14
Q8	.97***	.010	99.20
Q9	.90***	.028	32.32
Q10	.99***	.007	145.61
Q11	.99***	.005	218.90
Q12	1.0***	.004	256.88
Impacts (BF2) by			
Q13	.96***	.020	48.87
Q14	.97***	.010	94.70
Q15	.97***	.011	86.21
Q16	.75***	.059	12.80
Q17	.77***	.061	12.59
Q18	.99***	.010	102.43
Q19	.63***	.078	8.02
Q20	.67***	.078	8.53
Q21	.98***	.014	71.51
Q22	.95***	.017	54.84

Table 10 (Cont'd.)
Multilevel CFA Standardized (STDYX) Estimates for Within and Between Level

Factor/variable	Estimate	S.E.	Est./S.E.
Q23	$.99^{***}$	.017	58.60
Q24	1.0^{***}	.005	218.38
Q25	1.0^{***}	.004	232.14
Q26	$-.79^{***}$	.055	-14.32
Impacts (BF2) with			
Prerequisites (BF1)	$.88^{***}$	.023	37.75

Note. n=12,292. STDYX Standardization is based on background and outcome variables. "By" is short for "measured by" and is used to indicate the regression estimate between the underlying latent factors (F1-F8) and the observed indictor variables. "With" is short for "correlated with" and is used to indicate covariance relations between latent variables in the measurement model. Model includes 10 crossloadings. Complete indicator text is available in Table 1. **p* < .05. ***p* < .01. ****p* < .001.

Consistent with the standardized parameter estimates, the R^2 values are higher at the school level than the individual level (see Table 11). These values indicate the strength of each item in measuring its factor (Bryne, 2012). In the within model, evaluation process leading to better instruction (Q24) and evaluation process leading to improved student learning (Q25) are the two highest values (.92 and .91), whereas in the between model, multiple items have values above .9 (items above .9 are shaded in Table 6). Understanding of how measures are combined (Q3), feedback including guidance and/or suggestions for improvement (Q12), evaluation process leading to better instruction (Q27), and evaluation process leading to improved student learning (Q28) all have values above .99.

Table 11
R-Square Values for Within- and Between-Level

Variable	Estimate	S.E.	Est./S.E.
Within-level			
Q1	.61***	.010	60.07
Q2	.81***	.007	122.21
Q3	.87***	.006	137.39
Q4	.60***	.011	53.54
Q5	.31***	.013	23.99
Q6	.31***	.012	25.18
Q7	.80***	.007	122.55
Q8	.77***	.009	88.57
Q9	.65***	.013	51.66
Q10	.79***	.009	90.19
Q11	.83***	.007	126.89
Q12	.84***	.007	118.35
Q13	.65***	.010	67.02
Q14	.74***	.009	83.04
Q15	.71***	.010	71.55
Q16	.30***	.013	22.64
Q17	.32***	.013	24.80
Q18	.73***	.014	52.81
Q19	.20***	.011	17.45
Q20	.22***	.012	18.95
Q21	.80***	.010	84.33
Q22	.74***	.011	69.54
Q23	.70***	.011	64.47
Q24	.92***	.004	227.01
Q25	.91***	.005	186.11
Q26	.08***	.008	10.89
Between-level			
Q1	.80***	.051	15.61
Q2	.99***	.018	55.57
Q3	1.0***	.021	47.00
Q4	.66***	.061	10.82

Table 11 (Cont'd.)
R-Square Values for Within- and Between-Level

Variable	Estimate	S.E.	Est./S.E.
Q5	.49***	.080	6.10
Q6	.50***	.083	5.98
Q7	.94***	.020	48.07
Q8	.94***	.019	49.60
Q9	.81***	.050	16.16
Q10	.97***	.013	72.81
Q11	.98***	.009	109.45
Q12	.99***	.008	128.44
Q13	.92***	.038	24.44
Q14	.95***	.020	47.35
Q15	.94***	.022	43.10
Q16	.57***	.088	6.40
Q17	.60***	.095	6.29
Q18	.98***	.019	51.22
Q19	.40***	.099	4.01
Q20	.45***	.105	4.27
Q21	.96***	.027	35.75
Q22	.90***	.033	27.42
Q23	.97***	.033	29.30
Q24	1.0***	.009	109.19
Q25	.99***	.009	116.07
Q26	.63***	.087	7.16

Note. n=12,292.
**p* < .05. ***p* < .01. ****p* < .001.

The results of the multilevel model indicate that while an eight-factor model is a reasonable representation of teacher perceptions of a PES, at the group level, a broader two-factor model explains variation between schools. That is, it is a model with perceptions of understanding, fairness, evaluator credibility, and quality of feedback represented by one factor, and all of the impacts and benefits (PD, teaching, collegiality, and benefits) represented by another factor.

Evaluation Climate Strength Assessment

To identify those schools with a strong evaluation climate, we developed two simple composite indicators of the schools' average level of teachers' perceptions of the evaluation system and the degree of consensus around that level. Schools with a strong evaluation climate should have both a high level of positive perceptions about the evaluation system, and a high consensus (low mean absolute deviation) around the school average. Because we found that only two factors could be distinguished in the data at the school level, we created factor-based scores for each of the two and averaged them to get a composite indicator of climate level. We justified this combination based on the high correlation between the factors (.88) and the high correlation between factor-based scores (.77). Because the multilevel CFA supported the existence of the eight individual-level factors, we continued to work with the eight factor-based scales discussed above and we combined these by averaging across schools as well. This provided a simple way to depict climate strength for any school at the intersection of the level and consensus measures on a plot, such as Figure 4 below.

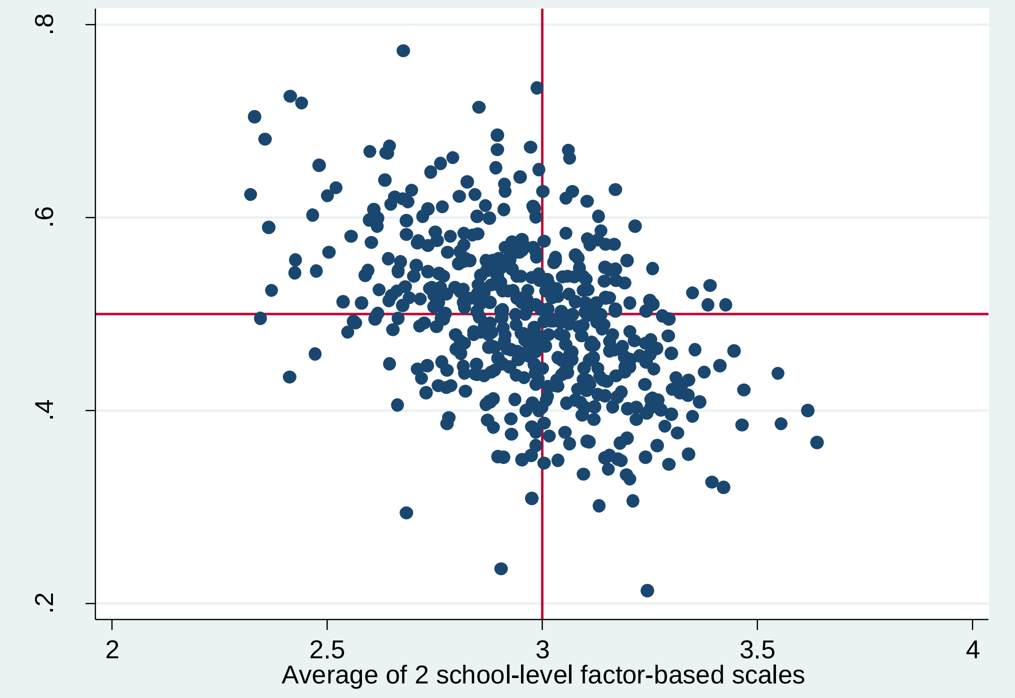

Figure 4. School Evaluation Climate Strength Based on Level and Consensus
Note. For schools with five or more respondents, *n*=496.

The schools in the lower right quadrant are plausibly the ones with the strongest climates. Schools to the right of the vertical line at a value of 3 have average responses to the items about the evaluation system that are favorable (3 was the scale point for agree or "to some extent" responses). Schools below the horizontal line have a mean absolute deviation below 0.5, a value chosen because it represents one-half of a scale point. There are 155 schools in this quadrant. Of course, the criteria for level and consensus could be set at different points. Moving the requirement for the level to 3.25 or above and the mean absolute deviation to below 0.4 identifies only 17 schools (about 3%) as having a strong climate. It is also likely that since this evaluation system was relatively new when the survey was administered, it is unrealistic to expect a strong climate to have developed in many schools.

Discussion

Dimensions of and Interrelationships Among Teachers' Perceptions of a Performance Evaluation System

This study sought to examine teacher perceptions of a PES to highlight different levers to improve the organizational effects evaluation systems. First, based on the argument that evaluatees' perceptions of the evaluation systems affect whether these systems will have the positive impacts on performance desired, we provided some further evidence that indicates teachers' perceptions of the evaluation components are interrelated, and that they are linked to perceptions of changes in practice and potential impact on student learning. This study contributes to the research and literature on conceptualizing and measuring teacher perceptions of a PES (e.g., Cherasaro et al., 2016; Jiang et al., 2015) by providing additional measurement scales and a structural model that demonstrates the structural relationships between multiple factors. Consistent with the theory of action and previous research (e.g., Cherasaro et al., 2016; Heneman & Milanowski, 2003; Williams & Levy, 2000) that suggests evaluatees' perceptions of the evaluation system are related to impacts on performance, we find that teachers' understanding the new evaluation system, perceptions of measures fairness, perceptions of evaluator credibility, the quality of feedback, impact on professional development, impact on collegiality, and impact on instruction are all structurally related.

These findings underscore the importance of examining teacher reactions to evaluation systems in formative and summative evaluations to identify any potential pitfalls in implementation of new PESs. Program administrators can use this type of information to examine whether the links in the theory of action are being connected and, if not, to identify areas where implementation of the evaluation system may need to be improved (e.g., communication of standards and procedures). To assess the relationship between teacher perceptions and actual changes in teaching practices and impact on student learning, further research could incorporate other measures (e.g., observation ratings, value-added measures) to examine the structural links between teacher perceptions of changes to instruction, actual changes in professional practice, and student learning.

Within- and Between-level School Variability

Furthermore, we argued that the multilevel nature of teacher perceptions should be examined to avoid faulty inferences about policy recommendations for PESs. In applied education research and policy literature, cross-level invariance is often assumed rather than assessed. When individual perceptions are aggregated to form group-level variables, it is assumed there is cross-level invariance in the measurement model between the individual (within-level) and group level (between-group). While making this assumption is intuitive and appealing, the results of the multilevel analysis demonstrate that in this case, cross-level invariance is not a tenable assumption. We found evidence that suggests there is cross-level measurement noninvariance, with fewer factors being identified at the school level. This finding indicates that it may not be appropriate for researchers and practitioners simply to aggregate teacher-level perceptions of PES to form school-level PES factors and draw inferences about the school. A between-level analysis may help one avoid making an individualistic fallacy, in which relationships between phenomena at the individual level are assumed to carry over to the group level (conversely, assessing school-level agreement levels helps one avoid making an ecological fallacy, in which individual perceptions are assumed to be well represented by the group average when there is substantial variation in individual perceptions). In addition to assisting researchers and practitioners in avoiding erroneous assumptions about cross-level noninvariance, this multilevel analysis suggests a need for a theory of action of PESs at the school level with corresponding school-level constructs.

Climate Strength: School-Level Agreement

Finally, we argued that one way for evaluation systems to have a positive effect on performance is to create a strong climate characterized by high levels of favorable perceptions along with high levels of agreement within schools. While we found little evidence that a shared strong evaluation climate had yet developed in this particular setting, we did illustrate some of the analyses that program administrators and evaluators could conduct to assess whether a strong climate is developing. The reason for the lack of a strong evaluation climate could be because this was the first year that all the district's teachers participated in the evaluation system. Unfortunately, we do not have any additional years of data to see if either the school averages changed or the consensus within schools improved. Nonetheless, at the early stage of implementation of the system we studied, we did find that for *Evaluator Credibility* and *Feedback Quality*, both positive average perceptions and relatively strong consensus had developed in more than a few schools. Teachers' perceptions of *Evaluator Credibility* and *Feedback Quality* were the most favorable aspects of the evaluation climate and the only two scales for which both positive average perceptions and relatively strong consensus had developed in more than a few schools. One reason these two stand out could be the effort the district had expended to train evaluators, coupled with the fact that many schools had only one evaluator. In contrast, perceptions of the fairness of the performance measures, which were chosen at the district level, is probably not something the school can influence as readily. Influencing teachers' understanding of the student growth measures may have been largely left to the district and teachers' association. Because of the multiple ways in which student achievement growth could be measured and combined with practice ratings, they may have been hard for principals to explain, as well as varying across teachers.

As expected, school size was negatively related to teacher perceptual consensus for all but one scale, and schools with higher average teacher-principal trust and perceived principal instructional leadership had slightly better agreement on the evaluator credibility scale, the one most likely to be influenced by the principal. Schools with higher average levels of teacher-principal trust, principal instructional leadership, program coherence, and teacher influence also had less consensus (higher average mean absolute deviations) on the benefit of the evaluation process. Though the relationship was small, this was still puzzling, since one might expect that more positive school leadership would foster more consensus. Unfortunately, teacher-level responses to items making up these scales were not available, so we could not examine the relationship at the individual level. These analyses did not uncover any other strong predictors of within-school agreement.

The results of these analyses illustrate the potential importance of examining within-school agreement, both to assess the reliability of between-school differences in average teacher perceptions, and to assess whether schools are developing the strong evaluation climate that is likely to be important in building a shared conception of good teaching. In particular, it may be useful to know if, in schools where perceptions are neither highly favorable or unfavorable, this is due to disagreement within the school or whether most teachers simply have middle-of-the-scale perceptions. Where perceptions are mildly unfavorable, the actions needed to improve teacher perceptions could be quite different in a school where most teachers had similar perceptions compared to one in which there was a bimodal distribution of perceptions, with most teachers having either highly unfavorable or favorable perceptions. It is possible that some schools could have evaluation "subclimates" that bear investigation. System administrators or program evaluators could examine the distributions of scale values within each school graphically. Techniques for identifying modality (see Xu et al., 2014) could also be used by more sophisticated analysts to screen a large number of schools to identify those with bimodal distributions.

In examining the strength of the evaluation climate, future researchers could also collect data on other potential aspects of climate strength, such as the perceived consequences of high or low performance, whether professional development opportunities that could help teachers achieve higher ratings on the performance measures were available, and the extent to which teachers interacted around the performance measures. A study designed to assess system strength would also include some direct measures of teachers' perceptions as to whether a strong climate existed in the school. Prior research on climate (e.g., Gonzalez-Roma et al., 2002) has often used items that directly ask about climate by making the referent of the item the group (e.g., "In my school, performance expectations are clear to teachers" or "In my school, teachers agree on how the rubrics are interpreted") rather than the individual. Group members' agreement about characteristics of a common referent is used as the consensus measure instead of agreement in perceptions of individuals' experience or situations. Both kinds of items can be useful in assessing climate strength because both shared individual experiences and perceptions of similarity of individual experience can be indicators.

Another limitation of the climate strength analysis is that no information on the consequences of a strong climate was available. We argued that an important consequence would be a shared conception of good teaching, but the survey did not contain a direct measure of whether teachers perceived a common conception or vision. And since the purpose of achieving a shared conception is to support a culture of high performance, it would also have been useful to have a measure of school performance.[4]

Despite these limitations, this article illustrates how teachers' responses to surveys about their perceptions of evaluation systems can be used to check whether some of the links in the theory of action connecting the evaluation process to improved instruction are being made. Here, this was done by estimating a structural equation model connecting scales based on individual teachers' perceptions related to key constructs derived from a theory of action. The article also illustrates the analysis of variation in perceptions across the individual and school level, and the degree to which teachers within a school agree in their perceptions. Calculating intra-class correlations shows how much differences in perceptions across the whole system are likely to be due to common conditions within schools, or whether responses are more likely to be influenced by individual teachers' characteristics and idiosyncratic experiences within the school. We found most of the variation was within school. The article also illustrated using mean absolute deviations to assess how well teachers within schools agreed in their perceptions, and found that agreement was not very high in most schools and was not highly related to many of the school characteristics we expected would influence agreement. This also supports the interpretation that teacher perceptions were not strongly influenced by common school conditions. Since a relatively small proportion of the variation in perceptions is common to schools, we did a multi-level confirmatory factor analysis to see if the factor structure was the same for schools and for individuals. A very different factor structure, which we found, suggested that it was not reasonable to compare the level of perceptions on the eight factors originally postulated (and confirmed by the individual level part of the multi-level CFA) on all eight factors. We thus used the factors from the school-level model to develop an evaluation climate level measure for schools. We also used the mean absolute deviations to develop a measure of consensus in perceptions of the level of key evaluation system characteristics, and we used the climate level and consensus measures to identify schools that were developing a stronger evaluation climate. Overall, the results suggest that to improve the evaluation climate, program administrators would need to address individual teachers' concerns about the process, perhaps beginning with focus groups or other qualitative data collection to understand why teachers' perceptions differ within schools.

[4] Unfortunately, our data sharing agreement with the survey administrator precluded sharing codes that would have allowed us to assess the relationships between the school-level scale averages and average mean absolute deviations and school-level value-added estimates.

References

Aksoy, E., & Bayazit, M. (2014). The relationships between MBO system strength and goal-climate quality and strength. *Human Resource Management, 53*(4), 505–525. https://doi.org/10.1002/hrm.21603

Bentler, P. M. (1990). Comparative fit indexes in structural models. *Psychological Bulletin*, *107*, 238–246. https://doi.org/10.1037/0033-2909.107.2.238

Bliese, P. D. (2000). Within-group agreement, nonindependence, and reliability: Implications for data aggregation and analyses. In K. J. Klein & S. W. J. Kozlowski (Eds.), *Multilevel theory, research, and methods in organizations: Foundations, extensions, and new directions* (pp. 349–381). San Francisco, CA: Jossey-Bass.

Bowen, D. E., & Ostroff, C. (2004). Understanding HRM-firm performance linkages: The role of the "strength "of the HRM system. *The Academy of Management Review, 29*(2), 203–221.

Bowen, D. E., & Schneider, B. (2014). A service climate synthesis and future research agenda. *Journal of Service Research, 17*(1), 5–22. https://doi.org/10.1177/1094670513491633

Browne, M. W., & Cudeck, R. (1993). Alternative ways of assessing model fit. In K. A. Bollen & J. S. Long (Eds.), *Testing structural equation models* (pp. 136–162). Newbury Park, CA: Sage.

Byrne, B. M.. (2012). *Structural equation modeling with Mplus: Basic concepts, applications and programming.* New York, NY: Routledge.

Cherasaro, T. L., Brodersen, R. M., Reale, M. L., & Yanoski, D. C. (2016). *Teachers' responses to feedback from evaluators: What feedback characteristics matter?* (REL 2017–190). Washington, DC: U.S. Department of Education, Institute of Education Sciences, National Center for Education Evaluation and Regional Assistance, Regional Educational Laboratory Central. Retrieved from http://ies.ed.gov/ncee/edlabs.

Coggshall, J. G., Rasmussen, C., Colton, A., Milton, J., & Jacques, C. (2012). *Generating teaching effectiveness: The role of job-embedded professional learning in teacher evaluation*: *A Research & Policy Brief.* Washington, DC: National Comprehensive Center for Teacher Quality. http://eric.ed.gov/?id=ED532776

Cole, M. S., Bedeian, A. G., Hirschfield, R. R., & Vogel, B. (2011). Dispersion-composition models in multi-level research: A data-analytic framework. *Organizational Research Methods*, *14*(4), 718–734. https://doi.org/10.1177/1094428110389078

Cronbach, L. J., & Meehl, P. E. (1955). Construct validity in psychological tests. *Psychological Bulletin*, *52,* 281–302. https://doi.org/10.1037/h0040957

Danielson, C. (2007). *Enhancing professional practice: A framework for teaching* (2nd ed). Alexandria, VA: Association for Supervision and Curriculum Development.

Darling-Hammond, L. (2013). *Getting teacher evaluation right: What really matters for effectiveness and improvement.* New York, NY, and Oxford, OH: Teachers College Press and Learning Forward.

Datnow, A., & Catellano, M. (2000). Teachers' responses to Success for All: How beliefs, experiences, and adaptions shape implementation. *American Educational Research Journal*, *37*(3), 775–799. https://doi.org/10.3102/00028312037003775

Desimone, L. M., Porter, A. C., Garet, M. S., Yoon, K. S., & Birman, B. F. (2002). Effects of professional development on teachers' instruction: Results from a three-year longitudinal study. *Educational Evaluation and Policy Analysis*, *24*(2), 81–112. https://doi.org/10.3102/01623737024002081

Dickson, M. W., Resick, C. J., & Hanges, P. J. (2006). When organizational climate is unambiguous, it is also strong. *Journal of Applied Psychology, 91*(2), 351–364. https://doi.org/10.1037/0021-9010.91.2.351

Firestone, W. A., Nordin, T. L., Shcherbakov, A., Kirova, D., & Blitz, C. L. (2014). *New Jersey's Pilot Teacher Evaluation Program: Year 2 final report.* New Brunswick, NJ: Rutgers Graduate School of Education.

Gitlin, A., & Margonis, F. (1995). The political aspect of reform: Teacher resistance as good sense. *American Journal of Education*, *103*(3), 377–405. https://doi.org/10.1086/444108

Gonzalez-Roma, V., Peiro, J. M., & Todera, N. (2002). An examination of the antecedents and moderator influences of climate strength. *Journal of Applied Psychology*, *87*(3), 465–473. https://doi.org/10.1037/0021-9010.87.3.465

Griffin, M. A., & Curcuruto, M. (2016). Safety climate in organizations. *Annual Review of Organizational Psychology and Organizational Behavior, 3*, 191–212. https://doi.org/10.1146/annurev-orgpsych-041015-062414

Halverson, R. R., & Clifford, M. A. (2006). Evaluation in the wild: A distributed cognition perspective on teacher assessment. *Educational Administration Quarterly, 42*(4), 578–619. https://doi.org/10.1177/0013161X05285986

Halverson, R. R., Kelley, C., & Kimball, S. (2004). Implementing teacher evaluation systems: How principals make sense of complex artifacts to shape local instructional practice. In W. K. Hoy & C. Miskel (Eds.), *Educational Administration, Policy, and Reform: Research and Measurement. A Volume in: Research and Theory in Educational Administration* (pp. 153–188). Information Age Publishing.

Hanushek, E. (2009). Teacher deselection. In D. Goldhaber & J. Hannaway (Eds.), *Creating a new teaching profession* (pp. 165–180). Washington, DC: Urban Institute Press.

Hedge, J. W., & Teachout, M. S. (2000). Exploring the concept of acceptability as a criterion for evaluating performance measures. *Group & Organization Management, 25*(1), 22–44. https://doi.org/10.1177/1059601100251003

Heneman, H. G. III, & Milanowski, A. T. (2003). Continuing assessment of teacher reactions to a standards-based teacher evaluation system. *Journal of Personnel Evaluation in Education, 17*(3), 171–195. https://doi.org/10.1023/B:PEEV.0000032427.99952.02

Henry, G. T., & Guthrie, E. (2015). *An evaluation of the North Carolina educator evaluation system and the student achievement growth standard 2010-11 through 2013-14.* Consortium for Educational Research and Evaluation–North Carolina.

Hu, L.-T., & Bentler, P. M. (1995). Evaluating model fit. In R. H. Hoyle (Ed.), *Structural equation modeling: Concepts, issues, and applications* (pp. 76–99). Thousand Oaks, CA: Sage.

Hu, L.-T., & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling, 6*, 1–55. https://doi.org/10.1080/10705519909540118

Ilgen, D. R., Fisher, C. D., & Taylor, M. S. (1979). Consequences of individual feedback on behavior in organizations. *Journal of Applied Psychology*, *64*(4), 349–371. https://doi.org/10.1037/0021-9010.64.4.349

Jiang, J. Y., Sporte, S. E., & Luppescu, S. (2015). Teacher perspectives on evaluation reform: Chicago's REACH students. *Educational Researcher*, *44*(2), 105–116. https://doi.org/10.3102/0013189X15575517

Kaplan, D. (2000). *Structural equation modeling: foundations and extensions*. Thousand Oaks, CA: Sage.

Kavanagh, P. Benson, J., & Brown, M. (2007). Understanding performance appraisal fairness. *Asia Pacific Journal Of Human Resources, 45*(2), 132–150. https://doi.org/10.1177/1038411107079108

Kinicki, A. J., Prussia, G. E., Wu, B., & McKee-Ryan, F. M. (2004). A covariance structure analysis of employees' response to performance feedback. *Journal of Applied Psychology*, *89*(6), 1057–1069. https://doi.org/10.1037/0021-9010.89.6.1057

Klugman, J., Gordon, M. F., Sebring, P. B., & Sporte, S. E. (2015). *A first look at the 5 Essentials in Illinois schools.* Chicago, IL: The University of Chicago Consortium on Chicago School Research.

Lindell, M. K., & Brandt, C. J. (2000). Climate quality and climate consensus as mediators of the relationship between organizational antecedents and outcomes. *Journal of Applied Psychology, 85*(3), 331–348. https://doi.org/10.1037/0021-9010.85.3.331

Longford, N. T., & Muthén, B. O. (1992). Factor analysis for clustered observations. *Psychometrika*, *57*, 581–597. https://doi.org/10.1007/BF02294421

Milanowski, A. T., & Heneman III, H. G. (2001). Assessment of teacher reactions to a standards-based teacher evaluation system: A pilot study. *Journal of Personnel Evaluation in Education, 15(3),* 193–212. https://doi.org/10.1023/A:1012752725765

Mischel, W. (1973). Toward a cognitive social learning reconceptualization of personality. *Psychological Review, 80,* 252–283. https://doi.org/10.1037/h0035002

Monk, D. H., & King, J. (1994). Multi-level teacher resource effects on pupil performance in secondary mathematics and science: The role of teacher subject matter preparation. In R. Ehrenberg (Ed.), *Choices and consequences: Contemporary policy issues in education* (pp. 29–58). Ithaca, NY: ILR Press.

Muthén, B. O. (1994). Multilevel factor analysis of class and student achievement components. *Journal of Educational Measuremen*t, *28*, 338–354. https://doi.org/10.1111/j.1745-3984.1991.tb00363.x

Muthén, B. O. (1997). Latent variable modeling of longitudinal and multilevel data. In A. E. Raftery (Ed.), *Sociological methodology 1997* (pp. 453–481). Washington, DC: American Sociological Association. https://doi.org/10.1111/1467-9531.271034

Muthén B. O., & Muthén, L. K. (2015). Mplus (Version 7.4) [Computer software]. Los Angeles, CA: Author.

O'Pry, S.C., & Schumacher, G. (2012). New teachers' perceptions of a standards-based performance appraisal system. *Educational Assessment, Evaluation and Accountability, 24*(4), 325–350. https://doi.org/10.1007/s11092-012-9148-4

Papay, J. (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

Reise, R. P., Ventura, J., Nuechterlein, K. H., & Kim, K. H. (2005). An illustration of multilevel factor analysis. *Journal of Personality Assessment*, *84*, 126–136. https://doi.org/10.1207/s15327752jpa8402_02

Schneider, B., Salvaggio, A. N., & Subirats, M. (2002). Climate strength: A new direction for climate research. *Journal of Applied Psychology, 87*(2), 220–229. https://doi.org/10.1037/0021-9010.87.2.220

Schweig, J. (2014). Cross-level measurement invariance in school and classroom environment surveys: Implications for policy and practice. *Educational Evaluation and Policy Analysis*, *36*(3), 259–280. https://doi.org/10.3102/0162373713509880

Schweig, J. D. (2016). Moving beyond means: Revealing features of the learning environment by investigating the consensus among student ratings. *Learning Environments Research*, *19*(3), 441–462. https://doi.org/10.1007/s10984-016-9216-7

Selig, J. P., Card, N. A., & Little, T. D. (2008). Latent variable structural equation modeling in cross-cultural research: Multigroup and multilevel approaches. In F. J. R. van de Vijver, D. A. Hmert, & Y. H. Poortinga (Eds.), *Multilevel analysis of individuals and cultures* (pp. 93–119). Mahwah, NJ: Erlbaum.

Sporte, S. E., & Jiang, J. Y. (2016). *Teacher evaluation in practice: Year 3 teacher and administrator perceptions of REACH.* [Research brief.] Chicago, IL: & University of Chicago Consortium on Chicago School Research.

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), 340–359. https://doi.org/10.1162/EDFP_a_00186

Tennessee Department of Education. (2016). Teacher and administrator evaluation in Tennessee: A report on year 4 implementation. Available at http://team-tn.org/wp-content/uploads/2013/08/TEAM-Year-4-Report1.pdf

Vianen, A. V., DePater, I. E., Bechtoldt, M. N., & Evers, A. (2011). The strength and quality of climate perceptions. *Journal of Managerial Psychology, 26*(1), 77–92. https://doi.org/10.1108/02683941111099637

Williams, J. R., & Levy, P. E. (2000). Investigating some neglected criteria: The influence of organizational level and perceived system knowledge on appraisal reactions. *Journal of Business and Psychology, 14*(3), 501–513. https://doi.org/10.1023/A:1022988402319

Xu, L., Bedrick, E. J., Hanson, T., & Restrepo, C. (2014). A comparison of statistical tools for identifying modality in body mass distributions. *Journal of Data Science, 12,* 175-196.

Appendix A

Table for Structural Equation Model Section

Table A1

Standardized (STDYX) Estimates for SEM (Model 2)

Factor	Estimate	*SE.*	Est./*SE*
F2 on			
F1	.59***	.007	81.39
F3 on			
F2	.63***	.007	87.72
F5 on			
F2	.90***	.004	228.06
F4 on			
F2	.35***	.010	35.86
F3	.55***	.009	61.44
F6 on			
F4	.18***	.010	17.89
F5	.85***	.010	88.25
F7 on			
F2	.86***	.006	153.45
F8 on			
F6	.22***	.015	14.72
F7	.62***	.015	42.48

Note: n=12,292. "On" is short for "regressed on" and is used to indicate the regression paths between latent factors. Model includes 10 crossloadings. Complete indicator text is available in Table 1.

$*p < .05$. $**p < .01$. $***p < .001$.

Appendix B

Distributions of Mean Absolute Deviations of Teacher Perception Scales

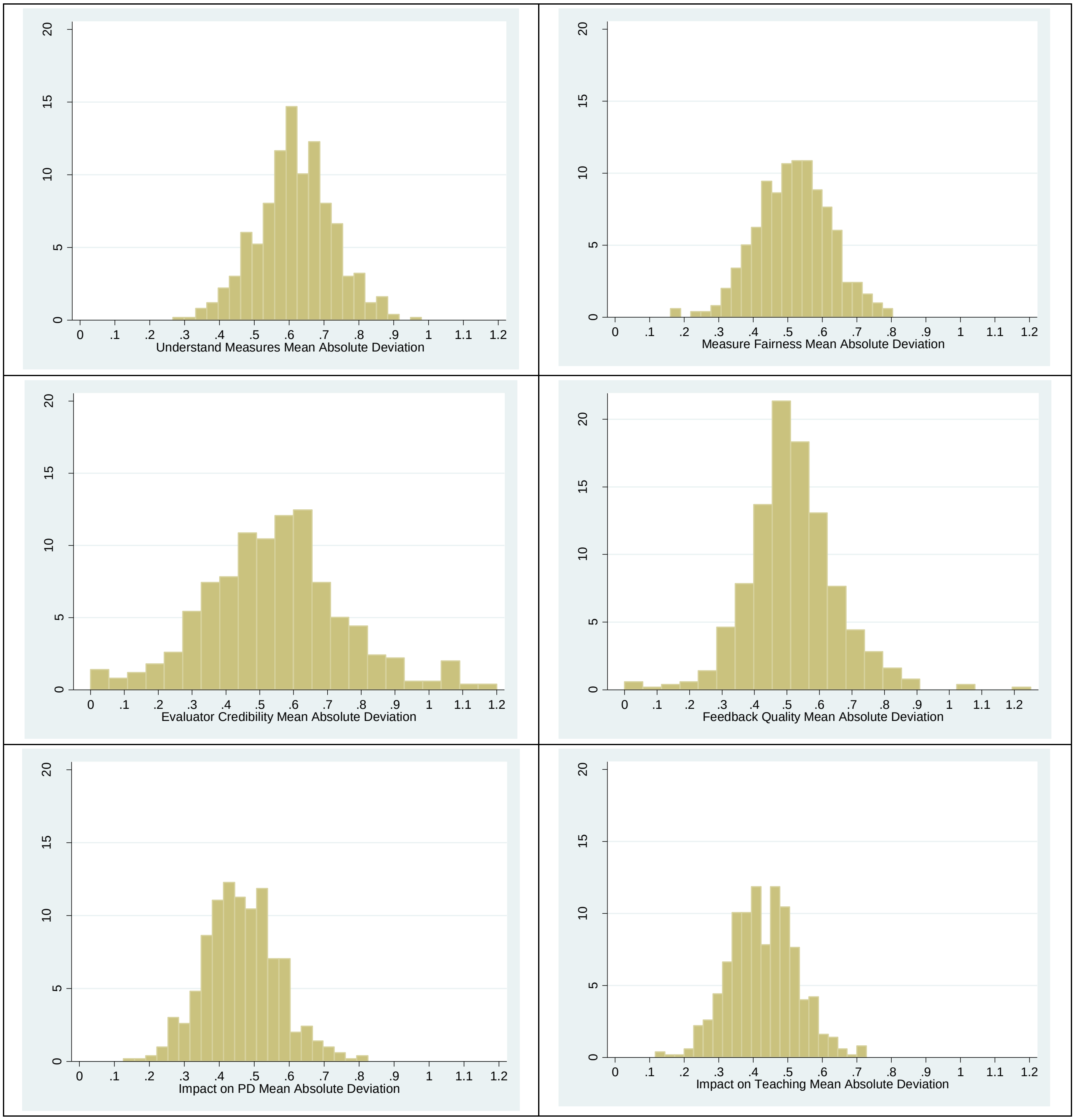

Figure B1. Distributions of Mean Absolute Deviations of Teacher Perception Scales
Note: n=496 schools with five or more respondents.

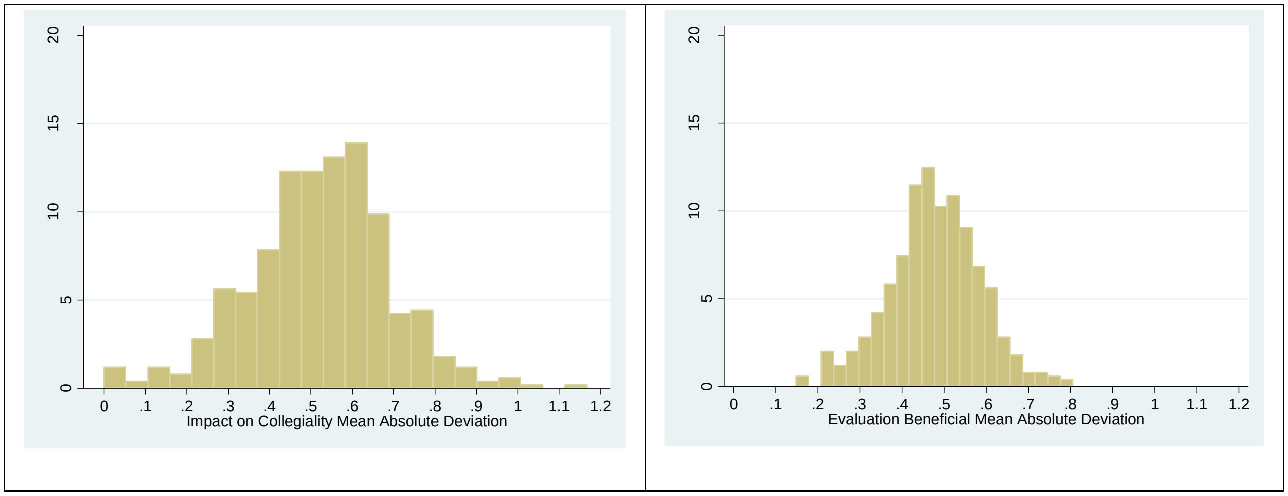

Figure B1 cont. Distributions of Mean Absolute Deviations of Teacher Perception Scales

Note: n=496 schools with 5 or more respondents.

Appendix C

School Leadership Scales – Example Items

Teacher Influence Example Items:

How much influence do teachers have over school policy in each of the areas below:

Hiring new professional personnel.

Setting standards for student behavior.

Response Options: Not at All, A Little, Some, To a Great Extent

Principal Instructional Leadership Example Items

The principal at this school:

Participates in instructional planning with teams of teachers.

Communicates a clear vision for our school.

Response Options: Strongly Disagree, Disagree, Agree, Strongly Agree

Program Coherence

To what extent do you disagree or agree with the following:

Many special programs come and go at this school.

Once we start a new program, we follow up to make sure that it's working.

Response Options: Strongly Disagree, Disagree, Agree, Strongly Agree

Teacher-Principal Trust

Please mark the extent to which you disagree or agree with each of the following:

It's OK in this school to discuss feelings, worries, and frustrations with the principal.

The principal looks out for the personal welfare of the faculty members.

Response Options: Strongly Disagree, Disagree, Agree, Strongly Agree

Source: Klugman, J., Gordon, M.F., Sebring, P. B., & Sporte, S.E. (2015).

About the Authors

Matthew Finster
Westat
MatthewFinster@westat.com

Matthew Finster is a senior research associate with Westat focusing on teacher quality and human capital management issues in education. At Westat, he has provided technical assistance to Federal grantees, such as Teacher Incentive Fund (TIF) grantees, in the area of human capital management, including performance compensation, teacher retention, teacher leadership, and educator evaluation systems. He has written or coauthored briefs on performance incentives, teacher retention, teacher leadership, and teacher evaluation. Dr. Finster received a Ph.D. in Education from the University of Washington-Seattle.

Anthony Milanowski
Westat
AnthonyMilanowski@westat.com

Anthony Milanowski was previously a senior researcher at Westat when the research for this article was conducted, but is now a research scientist at Education Analytics. His research has involved educator performance evaluation and compensation, as well as the alignment of human resource management practices with district goals and strategies for improving student outcomes. He also provided technical assistance to grantees participating in the U.S. Department of Education's Teacher Incentive Fund (TIF). Before joining Westat, he was an assistant scientist with the Wisconsin Center for Education Research at the University of Wisconsin-Madison. Dr. Milanowski received a Ph.D. in Industrial Relations from the University of Wisconsin-Madison and has taught courses on a variety of human resource management topics.

education policy analysis archives

Volume 26 Number 41 April 2, 2018 ISSN 1068-2341

arquivos analíticos de políticas educativas conselho editorial