Chapter 6
Understanding (Dis)Agreement in Student Ratings of Teaching and the Quality of the Learning Environment

Jonathan D. Schweig and José Felipe Martínez

Abstract Student surveys are increasingly being used to collect information about important aspects of learning environments. Research shows that aggregate indicators from these surveys (e.g., school or classroom averages) are reliable and correlate with important climate indicators and with student outcomes. However, we know less about whether within-classroom or within-school variation in student survey responses may contain additional information about the learning environment beyond that conveyed by average indicators. This question is important in light of mounting evidence that the educational experiences of different students and student groups can vary, even within the same school or classroom, in terms of opportunities for participation, teacher expectations, or the quantity and quality of teacher–student interactions, among others. In this chapter, we offer an overview of literature from different fields examining consensus for constructing average indicators, and consider it alongside the key assumptions and consequences of measurement models and analytic methods commonly used to summarize student survey reports of instruction and learning environments. We also consider recent empirical evidence that variation in student survey responses within classrooms can reflect systematically different experiences related to features of the school or classroom, instructional practices, student background, or a combination of these, and that these differences can predict variation in important academic and social-emotional outcomes. In the final section, we discuss the implications for evaluation, policy, equity, and instructional improvement.

Keywords Student ratings · Learning environment · Teaching quality

J. D. Schweig (✉)
RAND Corporation, Arlington, VA, USA
e-mail: Jonathan_Schweig@rand.org

J. F. Martínez
University of California, Los Angeles, CA, USA
e-mail: jfmtz@ucla.edu

1 Introduction

Educators are increasingly turning to student surveys as a valuable source of information about important features of school and classroom learning environments, ranging from time on task and content coverage to more qualitative aspects of teaching—e.g., the extent to which classes are well-managed, teachers foster student cognitive engagement, or students feel emotionally, physically, and intellectually safe (Baumert et al., 2010; Klieme et al., 2009; Pianta & Hamre, 2009). Considerable research shows that student survey reports can be aggregated into reliable indicators of constructs that have been variously identified in the literature with terms like *learning environment*, *classroom climate*, *instructional practice*, or *teaching quality*. These constructs may or may not be exchangeable across areas of study, but irrespective of terminology, the literature shows that student survey aggregates tend to correlate significantly with each other, with indicators derived through other methods (e.g., classroom observation), and with a range of desirable student outcomes. However, there is a gap in research investigating whether within-classroom or within-school variability in such student survey responses may offer additional information beyond that conveyed by average indicators. This question is important in light of emerging evidence that the educational experiences of individual students can vary considerably within schools, and even within the same classroom, including opportunities for student participation (Reinholz & Shah, 2018; Schweig et al., 2020), and the quantity and quality of teacher–student interactions (e.g., Connor et al., 2009), among others. In this chapter we review literature that examines aggregate survey indicators in different fields, and consider the key assumptions and consequences of various measurement models and analytic methods commonly used to summarize student survey reports of teaching. We then examine the growing literature that investigates the variability in student survey responses within classrooms and schools, and whether this variation may relate to educational experiences and outcomes. We illustrate the potential implications of this kind of variation using a hypothetical example case. In the final section, we discuss the implications of this research for evaluation policy and instructional improvement.

2 Student Surveys, Teaching, and the Learning Environment

There are many reasons why educators are increasingly interested in student surveys as a source of information about learning environments. Perhaps most importantly, students can spend over 1,000 hours in their schools every year, and thus have unmatched depth and breadth of experience interacting with teachers and peers (Ferguson, 2012; Follman, 1992; Fraser, 2002). Students also provide a unique perspective compared to other reporters (Downer et al., 2015; Feldlaufer et al., 1988). Probing students about their perceptions of teaching and the learning environment

acknowledges their voice (Bijlsma et al., 2019; Lincoln, 1995), and the significance of their school-based experiences (Fraser, 2002; Mitra, 2007). Second, a growing body of research suggests that students can provide trustworthy information about important aspects of the learning environment (Marsh, 2007). For example, survey-based aggregate indicators can reliably distinguish among instructional practices (Fauth et al., 2014; Kyriakides, 2005; Wagner et al., 2013), and aspects of teaching quality (e.g., Benton & Cashin, 2012). These aggregates are furthermore significantly and positively associated with other measures of teaching quality (e.g., Burniske & Meibaum, 2012; Kane & Staiger, 2012).

Like other measures, student survey responses can be susceptible to error (e.g., recall, inconsistency in interpretation; see e.g., Popham, 2013; van der Lans et al., 2015), bias (e.g., acquiescence), and halo effects (perceptions of one aspect of teaching influencing those of other aspects; see e.g., Fauth et al., 2014; Chap. 3 by Röhl and Rollett of this volume) that may influence their psychometric properties (see for example, Follman, 1992; Schweig, 2014; Wallace et al., 2016). Neverthe-less, most existing studies suggest that these biases are generally small in magnitude and do not greatly influence comparisons across teachers or student groups, or how aggregates relate with one another and with external variables (Kane & Staiger, 2012; Vriesema & Gehlbach, 2019). Research also demonstrates that aggregated student survey responses are associated with important student outcomes including academic achievement (Durlak et al., 2011; Shindler et al., 2016), engagement (Christle et al., 2007), and self-efficacy and confidence (e.g., Fraser & McRobbie, 1995).

Student surveys also have the benefit of being cost-effective, relatively easy to administer, and feasible to use at scale (e.g., Balch, 2012; West et al., 2018). This is a particular advantage when contrasted with other commonly used methods for measuring teaching and the quality of the learning environment, including direct classroom observation. In large school districts, an observation system closely tied to professional development can require dozens of full-time positions, with yearly costs in the millions of dollars (Balch, 2012; Rothstein & Mathis, 2013). As a result, the use of student surveys has seen remarkable growth over the last two decades for evaluating educational interventions (Augustine et al., 2016; Gottfredson et al., 2005; Teh & Fraser, 1994), and monitoring and assessing educational programs and practices (Hamilton et al., 2019). In particular, student surveys are commonly used to inform teacher evaluation and accountability systems—summatively as input for setting actionable targets (Burniske & Meibaum, 2012; Little et al., 2009), or formatively to provide feedback and promote teacher reflection and instructional improvement (Bijlsma et al., 2019; Gehlbach et al., 2016; Wubbels & Brekelmans, 2005).

3 Psychological Climate, Organizational Climate, and Student Surveys

In most contexts, schooling is an inherently social activity, and students typically experience schooling in organizational clusters (Bardach et al., 2019). The common pattern of student clustering within classrooms and schools presents challenges and choices in using surveys to understand teaching and the quality of the learning environment. One of the first choices is whether to focus the survey on understanding the personal perceptions and experiences of individual class members, or more broadly on shared elements of teaching quality relevant to the class or school as a whole (Bliese & Halverson, 1998; Den Brok et al., 2006; Echterhoff et al., 2009).

Surveys that aim to capture individual student interpretations of teaching quality or of the learning environment are described as reflecting *psychological climate*, and include items that ask for individual self-perceptions and personal beliefs (Glick, 1985; Maehr & Midgley, 1991). A long history of educational research suggests that psychological climate is a key proximal determinant of academic beliefs, behaviors, and emotions (Maehr & Midgley, 1991; Ryan & Grolnick, 1986). Because psychological climate variables treat individual perceptions as interpretable, it is appropriate to analyze them at the individual level (Stapleton et al., 2016), and differences among individual respondents are considered as substantively meaningful. Individuals can react in different ways to the same practices, procedures that seem fair to one individual might seem unfair to another individual, and so forth. Psychological climate variables can be aggregated to describe the composition of an organization (Sirotnik, 1980).

On the other hand, surveys that focus on the classroom or the school as a whole are described as reflecting *organizational climate* (see e.g., Lüdtke et al., 2009; Marsh et al., 2012), a concept that has a rich history in industrial and social psychology (Bliese & Halverson, 1998; Chan, 1998). Unlike psychological climate, organizational climate emerges from the collective perceptions of individuals as they experience policies, practices, and procedures (e.g., Hoy, 1990; Ostroff et al., 2003). Aggregating individual perceptions produces measures of organizational level phenomena (Sirotnik, 1980). These new variables can be interpreted to reflect an *overall* or shared perception of the environment (Lüdtke et al., 2009). The concept of organizational climate informs the design and use of many student surveys, which are typically directed toward students as a group, often asking for observations of the behavior of others (e.g., classmates, teachers; see Den Brok et al., 2006).

When conceived as measures of organizational climate, aggregating survey responses essentially positions students as informants or judges of a classroom or school level trait, similar to observers who would provide ratings using a standardized protocol. To illustrate this assumption, consider the following claims in Table 1 regarding three widely used student surveys.

Thus, while psychological climate variables treat interindividual differences as substantively interpretable, organizational climate variables emerge based on shared student experiences, and assume that students have similar mental images of their

Table 1 Measurement claims for three widely used surveys

Survey	Claim
Tripod Survey	The variance between teachers provides the "signal" we are interested in… while the variability among students within a classroom may be regarded as "rater variance." In effect, Tripod casts each student within a class as an informant, or rater, of the quality of the classroom; inconsistencies among student responses within a class are therefore regarded as "rater error" and are thus part of the measurement error *Source: Raudenbush and Jean (2014), p. 179*
National Education Longitudinal Study	In the measurement sense, students are considered *judges* or raters of the disciplinary climate of the school. If the variation of ratings within schools is small, we consider inter-rater agreement to be strong *Source: Ma and Willms (2004), p. 174*
Learning Environment Scale	Ratings obtained by multiple informants within a structural class (e.g., multiple students within a school or multiple teachers within a school) can be considered interchangeable because they share a more common role and a presumed more similar perspective than informants from different structural groups *Source: Konold and Cornell (2015)*

classroom or school (Fraser, 1998). Students in a particular classroom or school are treated as *exchangeable* (Lüdtke et al., 2009), and interindividual differences are treated as idiosyncratic measurement error. Lüdtke et al. (2006, p. 207) noted that in the ideal scenario, "each student would assign the same rating, such that the responses of students in the same class would be interchangeable." Because organizational climate variables treat individual perceptions as error, it is appropriate to analyze them at the classroom or school level (Stapleton et al., 2016). However, while the distinction between psychological and organizational variables is frequently drawn in the theoretical and methodological literature, much-applied literature does not explicitly or consistently consider student survey-based ratings of teaching quality and the learning environment as either psychological or organizational level measures (Lam et al., 2015; Schweig, 2014; Sirotnik, 1980). This in part reflects the fact that most student surveys occupy a gray area between these two classifications. On one hand classrooms and schools are shared spaces, students interact socially and build social relationships with their peers and with their teachers, and some aspects of teaching quality are more or less equally applicable to all students in the classroom (Lam et al., 2015; Urdan & Schoenfelder, 2006). At the same time, students' school-based experiences can and often do differ, making their responses not exchangeable; students are not objective, external observers, but active participants involved in complex interactions with other students, teachers, and features of the classroom and school environment. Teachers often interact with students through multiple modes

and formats, both individually, and as a group (whole-class instruction, group work); and of course students interact directly with one another individually and as a group (Den Brok et al., 2006; Glick, 1985; Sirotnik, 1980).

4   Reporting Survey Results: Common Practices and Opportunities for Improvement

In the previous section, we argued that research often does not explicitly state the measurement assumptions that underlie their use of student surveys. In particular, researchers are not always explicit about the unit of interest (e.g., the individual or the group), and what this implies for the interpretability of individual student responses. These issues also arise in how survey developers choose to summarize and report survey results. In practice, nearly all survey platforms report measures of teaching and the quality of the learning environment by aggregating individual student responses to create classroom-level or school-level scores. It is these aggregates that are subsequently communicated to stakeholders or practitioners through data dashboards or survey reports (Bradshaw, 2017; Panorama Education, 2015). These aggregates can reflect simple averages (Balch, 2012; Bijlsma et al., 2019), percentages of respondents that report a certain experience or behavior (Panorama Education, 2015), or more sophisticated statistical models (e.g., IRT, or other latent variable models, see e.g., Maulana et al., 2014).

Irrespective of whether the survey developers are interested in individual or school- or classroom-level variables, this approach to score reporting often does not include information about the variability of student responses within classrooms or schools (Chan, 1998; Lüdtke et al., 2006). Thus, whether by accident or design, survey reports are ultimately firmly rooted in the notion of organizational climate in industrial or organizational psychology described previously: the shared learning environment is the central substantive focus, students are assumed to react similarly to similar external stimuli, and individual variation is assumed idiosyncratic or reflective of random measurement error (Chan, 1998; Lüdtke et al., 2009; Marsh et al., 2012).

However, while aggregated scores are useful for characterizing the overall learning experiences of a typical student, a growing body of research shows that these experiences can in fact vary greatly within schools and classrooms. Croninger and Valli (2009), for example, found that the vast majority (more than 80 percent) of the variance in the quality of spoken teacher–student exchanges occurred among lessons delivered by the same teachers. Den Brok et al. (2006) found that the majority of variance in student survey reports reflects differences among students within the same classroom (between 60 and 80 percent of the total variance). Crucially, emerging research also suggests that disagreement among students in their reports of the learning environment does not reflect only error, and indeed can provide important additional insights into teaching and learning, not captured by classroom or

school aggregates. In a study of elementary school students, Griffith (2000) found that schools with higher levels of agreement in student and parent survey reports of order and discipline tended to have higher levels of student achievement and parent engagement. Recent work by Bardach and colleagues (2019) found that within-classroom consensus on student reports of classroom goal structures was positively associated with socio-emotional and academic outcomes.

4.1   An Example Case of Within-Classroom Variability

Examining the distribution of student reports can open up possibilities for using information about the nature and extent of student disagreements for diagnostic and formative uses, and focused professional development opportunities for teachers, among others. The three hypothetical Classrooms in Fig. 1 illustrate how different within-classroom distributions can produce the same aggregate classroom climate rating (e.g., Lindell & Brandt, 2000; Lüdtke et al., 2006).

For the purposes of this example, students in each of these classrooms are asked about their perceptions of cognitive activation in the classroom, and the extent to which they are presented with questions that encourage them to think thoroughly and explain their thinking (Lipowski et al., 2009). Figure 1 displays the ratings provided by twenty students in each of the three classrooms. All three classrooms have the same average score of 3.42 on a 5-point scale.

In Classroom 1 there is noticeable disagreement in student survey responses, and students provide responses all across the allowable score range. In Classroom 2, there is also a lot of variability in student responses, but student perceptions seem polarized: there is a large group of students that feel very positively about the level
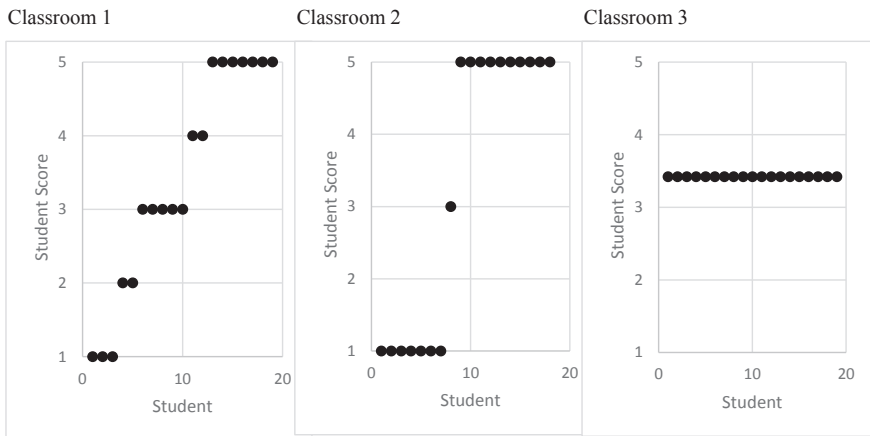

Fig. 1   Three hypothetical distributions of student climate ratings yielding the same average of 3.42

of cognitive activation, while a large group of students feel very negatively. Finally, in Classroom 3, there is perfect agreement among all students—this is the hypothetical ideal classroom described in Lüdtke and colleagues (2006) where all students experience classroom climate the same way. These scenarios raise important questions for practice. In principle it does not seem justifiable to give the three schools in Fig. 1 the same feedback and professional development recommendations for teachers—thus omitting the fact that the patterns of within-classroom variation are dramatically different. A more sensible approach would likely entail considering whether the within-classroom variability in student reports can potentially be informative for purposes of diagnosing and improving teaching quality. It is not possible to determine from this raw quantitative display *why* students in these three classrooms perceived cognitive activation in different ways. However, examining the distribution of student reports can open up possibilities for using this information for diagnostic and formative uses, and focused professional development opportunities for teachers. In the remainder of this chapter, we summarize and discuss relevant literature for understanding these interindividual differences.

5 School and Classroom Factors Associated with Variation in Student Perceptions of Teaching Quality

Within classrooms or schools, interindividual differences in the perception of teaching or the learning environment can arise for many reasons. We begin this section discussing the standard assumption invoked by common approaches to survey score reporting (that within-classroom or school variation reflects measurement error) and subsequently present four alternative interpretations that have support in the literature in other areas: (1) differential expectations and teacher treatment, (2) diversity of student needs and expectations, (3) diversity of student backgrounds, experiences, cultural values, and norms, and (4) teacher characteristics.

5.1 Measurement Error

Interindividual variability in student perceptions can be assumed to involve some idiosyncratic component of measurement error—i.e., random fluctuations around the "true" score of a school or classroom, related to memory, inconsistency, and unpredictable interactions among time, location, and personal factors. Individual students may also vary in terms of their standards of comparison (Heine et al., 2002), or the internal scales they use to calibrate their perceptions (Guion, 1973). This can create differences in student scores analogous to rater effects in studies of observational protocols: some students may be more lenient or severe than others. Thus, some differences among students are not substantively interpretable (Marsh et al., 2012;

Stapleton et al., 2016), Moreover, to the extent students are not systematically sorted into classrooms based on *stringency*, these differences are not expected to induce bias and are best treated as measurement error (West et al., 2018). If interindividual variability were idiosyncratic and random, however, we would generally not expect within-classroom student ratings to be associated with other measures of teaching quality or student outcomes. However, a number of prior studies have demonstrated that individual perceptions of school or classroom climate can be positively associated with student achievement. Griffith (2000) and Schweig (2016), found that learning environments with more intraindividual disagreements about order, discipline, and the quality of classroom management had lower academic performance, even holding average ratings constant. Schenke et al. (2018) found that lower levels of heterogeneity among students' perceptions of emotional support, autonomy support, and performance focus are negatively associated with mathematics achievement. Martínez (2012) found that individual perceptions of opportunity to learn (OTL) were predictive of reading achievement, even after controlling for class and school level OTL. Such findings strongly suggest that within-classroom variability in student reports is not entirely reflective of measurement error.

5.2   Differential Expectations and Teacher Treatment

Teacher expectations are a critical determinant of student learning (Muijs et al., 2014). Teachers may consciously or unconsciously have differential expectations for subgroups of students, which may translate into different sets of rules, classroom environments, and pedagogical strategies (Babad, 1993; Brophy & Good, 1974), potentially leading to opportunity gaps (Flores, 2007). Research has shown some teachers can have lower achievement expectations for students of color (Banks & Banks, 1995; Oakes, 1990). Teachers may also have lower achievement expectations for female students (Lazarides & Watt, 2015), and offer them less reinforcement and feedback (e.g., Simpson & Erickson, 1983). Teacher expectations may also differ based on perceptions of student ability. At higher grades, research has shown that prior academic achievement is the most significant influence on teacher expectations (Lockheed, 1976). More recent research suggests that learning tasks are often differentially assigned to students based on teacher beliefs about student ability. For example, "mathematically rich" instruction (tasks requiring reasoning and creativity, multiple concepts and methods, and application to novel contexts) is often reserved for students perceived to be *high-achieving*, while those perceived as lower achieving spend more time developing and practicing basic skills (Schweig et al., 2020; Stipek et al., 2001). Thus, within-classroom variability in student survey reports could point to suboptimal or inequitable participation opportunities and instructional experiences for students of different groups (Gamoran & Weinstein, 1998; Seidel, 2006), which may, in turn, result in achievement gaps (Voight et al., 2015).

5.3   Diversity of Student Needs and Expectations

Student perceptions of teaching and the learning environment may reflect different student needs and expectations—learning experiences and instructional practices that are successful with some students may not be effective with others, and student socio-emotional needs and expectations may also differ substantially within classrooms. Levy et al. (2003) provide an example that students with lower self-esteem may have greater needs with respect to the establishment of a supportive climate. Lüdtke et al. (2006) suggest that higher and lower ability students may differ in their perceptions of certain aspects of instructional practice, including pacing or task difficulty. English learners (ELs) and students with disabilities tend to report their schools to be less safe and supportive than their peers (Crosnoe, 2005; De Boer et al., 2013; Watkins & Melde, 2009). ELs face challenges with language comprehension, particularly with academic or mathematical language (Freeman & Crawford, 2008), and this may create differential perceptions of the clarity of classroom procedures. On the other hand, Hough and colleagues (2017) found that ELs had systematically more favorable perceptions of their teachers and classrooms than their peers on several aspects of climate. ELs students could be more engaged, more challenged, and better behaved, which influences their overall perception of the classroom (LeClair et al., 2009). In this way, ELs could also be more proactive at seeking out additional support from teachers, or that teachers are particularly sensitive to the needs of ELs (LeClair et al., 2009).

   Alternatively, teachers may use instructional strategies that are responsive to and supportive of students' diverse needs and expectations, potentially causing student perceptions of the quality of their learning experiences to be *more similar*. For example, teachers may use complex instruction structured to promote student engagement, support critical thinking, and to connect content in meaningful ways to students' lives (Averill et al., 2009; Freeman & Crawford, 2008). Thus, to the extent that within-classroom agreement is associated with the use of instructional strategies responsive to students' diverse needs and expectations, there may be more equitable opportunities for all students. In a recent mixed-methods study of science classrooms, we found that classrooms with higher levels of student agreement tended to provide more collaborative learning opportunities for students, including more group work, and to have more structured systems for eliciting student participation (Schweig et al., n.d.).

5.4   Diversity of Student Backgrounds, Experiences, Cultural Values, and Norms

Reports of teaching and the quality of the learning environment may reflect cultural or contextual factors that cause students to perceive the learning environment differently (Bankston & Zhou, 2002; West et al., 2018). There is also research suggesting that

student perceptions of the learning environment may also differ by grade level (West et al., 2018). In the United States, research has shown that Black and Hispanic/Latino students often report feeling less connected to their schools, feel less positively about their relationships with teachers and administrators, and feel less safe in some areas of the school (Lacoe, 2015; Voight et al., 2015). However, recent literature suggests that this may not always be the case. Hough and colleagues (2017) found that while Black students had systematically lower ratings of school connectedness, discipline, and safety than their peers, Hispanic/Latino students tended to report systematically higher perceptions. These findings are not inherently at odds, and other literature suggests that perceptions of the learning environment can differ even from one area of the school to the other. Using data from New York City, Lacoe (2015) found that, for example, Black students have systematically lower perceptions of safety than their white peers in classrooms, but have systematically higher perceptions of safety in hallways, bathrooms, and locker rooms. In our own work, we found that classes with higher proportions of ELs and low-achieving students tended to have more intraindividual disagreements about teaching and the quality of the learning environment (Schweig, 2016; Schweig et al., 2017) in mathematics and science classrooms, and we also found significant within-classroom gaps between Black and white students on several aspects of teaching and the quality of the learning environment, with Black students typically having more positive perceptions relative to their white peers (Perera & Schweig, 2019).

The perception of some teacher behaviors, including the extent to which teachers make students feel cared for, may depend strongly on cultural conceptions of caring (Garza, 2009). Calarco (2011) highlighted several ways in which economically disadvantaged students help-seeking behaviors differed from their classmates in ways that could impact perceptions of teaching quality. Specifically, Calarco found that economically disadvantaged students sought less teacher assistance, and as a result, received less guidance from their teachers. Atlay and colleagues (2019) found that students from higher socioeconomic backgrounds were more critical about teacher assistance, perhaps reflecting a sense of entitlement (Lareau, 2002). Students' perceptions of teaching quality can also be influenced by out-of-school experiences. For example, there may be differential exposure to external stressors that influence feelings of school safety (Bankston & Zhou, 2002; Lareau & Horvat, 1999).

5.5   Teacher Characteristics

A number of teacher characteristics can influence survey-based reports. Past work, for example, has shown that student perceptions of teachers are associated with teacher experience, and in particular, that more experienced teachers are perceived as more dominant and strict (Levy & Wubbels, 1992). More experienced teachers, however, are not generally perceived as more caring or supportive by their students (Den Brok et al., 2006; Levy et al., 2003). Teacher race and ethnicity can also play a role in survey-based ratings of teaching quality. Newly emerging research suggests

that race-based disparities in perceptions of teaching quality can be ameliorated by the presence of teachers of color. Specifically, teacher–student race congruence may positively influence students' perceptions of teaching quality (Dee, 2005; Gershenson et al., 2016). In our own research, however, we did not find evidence that observable teacher characteristics, including teacher race, gender, years of experience, and level of education explain variation in race-based perceptual gaps (Perera & Schweig, 2019; Schweig, 2016).

6 Conclusion

A growing body of evidence suggests that in considering instructional climate, researchers and school leaders may want to look beyond aggregate indicators, and consider also the extent of variation (or consensus) in student survey reports, as a potential indicator of important aspects of the school or classroom environment. In fact, the ability to capture within-school or within-classroom variability in student experiences is one of the defining strengths of student survey-based measures. Other commonly used measurement modes (including teacher self-report and structured classroom observations) are structurally not well-equipped to capture differential student experiences. Classroom observation protocols, for example, are typically not designed to measure whether or how teachers engage with individual students (Cohen & Goldhaber, 2016; Douglas, 2009). Student surveys, on the other hand, offer information that goes beyond typical experiences and can allow teachers and instructional leaders better understand how instruction, socio-emotional support, and other aspects of the learning environment are experienced by different students or groups of students.

Collectively, the research presented in this chapter suggests that variation in student survey reports of their learning environment may reflect a variety of factors and influences, ranging from strategic instructional choices, responsive pedagogy, and classroom structures implemented by teachers, varying needs and perceptions of particular students or groups of students, contextual factors, and the interactions among these. Importantly, variation can also reflect more pernicious influences like differential teacher expectations, and other structural disadvantages for some group of students. Our example case also raises important questions about whether within-school or within-classroom variability should be considered as ignorable measurement error when examining student survey reports of teaching quality and learning environments. Should we give the three classrooms in Fig. 1 the same feedback and professional development recommendations for teachers? Or is there evidence in the within-classroom variability in student reports that can potentially be informative for these purposes? Recent policy guidelines in the United States either explicitly require or implicitly move in the latter direction, advising education agencies to provide schools not only aggregated survey-based indicators, but also indicators disaggregated by student subgroup (Holahan & Batey, 2019; Voight et al., 2015). A growing consensus also sees attending to these subgroup differences as a key for

school-wide adoption of instructional improvement strategies that meet the learning needs of the most vulnerable students (Kostyo et al., 2018).

Considering the diversity of student perspectives and experiences can be particularly useful for informing efforts to promote equitable learning and outcomes. Ultimately, whether the climate is conceived as a psychological or organizational climate, or both, if subgroups of students experience school life in meaningfully different ways, reliance on aggregated survey indicators as measures of teaching quality can potentially obscure diagnostic information (Roberts et al., 1978), and compromise the validity and utility of these measures to inform teacher reflection or feedback, and other improvement processes within schools (Gehlbach, 2015; Lüdtke et al., 2006).

References

Atlay, C., Tieben, N., Fauth, B., & Hillmert, S. (2019). The role of socioeconomic background and prior achievement for students' perception of teacher support. *British Journal of Sociology of Education, 40*(7), 970–991. https://doi.org/10.1080/01425692.2019.1642737.

Augustine, C. H., McCombs, J. S., Pane, J. F., Schwartz, H. L., Schweig, J., McEachin, A., & Siler-Evans, K. (2016). *Learning from summer: Effects of voluntary summer learning programs on low-income urban youth.* RAND Corporation.

Averill, R., Anderson, D., Easton, H., Maro, P. T., Smith, D., & Hynds, A. (2009). Culturally responsive teaching of mathematics: Three models from linked studies. *Journal for Research in Mathematics Education*, 157–186.

Babad, E. (1993). Teachers' differential behavior. *Educational Psychology Review, 5*(4), 347–376. https://doi.org/10.1007/BF01320223.

Balch, R. T. (2012). *The validation of a student survey on teacher practice.* Doctoral dissertation, Vanderbilt University, Nashville, TN.

Banks, C. A., & Banks, J. A. (1995). Equity pedagogy: An essential component of multicultural education. *Theory Into Practice, 34*(3), 152–158. https://doi.org/10.1080/00405849509543674.

Bankston III, C. L., & Zhou, M. (2002). Being well vs. doing well: Self esteem and school performance among immigrant and nonimmigrant racial and ethnic groups. *International Migration Review, 36*(2), 389–415. https://doi.org/10.1111/j.1747-7379.2002.tb00086.x.

Bardach, L., Yanagida, T., Schober, B., & Lüftenegger, M. (2019). Students' and teachers' perceptions of goal structures–will they ever converge? Exploring changes in student-teacher agreement and reciprocal relations to self-concept and achievement. *Contemporary Educational Psychology, 59.* https://doi.org/10.1016/j.cedpsych.2019.101799.

Baumert, J., Kunter, M., Blum, W., Brunner, M., Voss, T., Jordan, A., Klusmann, U., Krauss, S., Neubrand, M., & Tsai, Y. M. (2010). Teachers' mathematical knowledge, cognitive activation in the classroom, and student progress. *American Educational Research Journal, 47*(1), 133–180. https://doi.org/10.3102/0002831209345157.

Benton, S. L., & Cashin, W. E. (2012). *Student ratings of teaching: A summary of research and literature.* IDEA Center. https://www.ideaedu.org/Portals/0/Uploads/Documents/IDEA%20Papers/IDEA%20Papers/PaperIDEA_50.pdf. Accessed 5 August 2020.

Bijlsma, H. J. E., Visscher, A. J., Dobbelaer, M. J., & Bernard, P. (2019). Does smartphone-assisted student feedback affect teachers' teaching quality? *Technology, Pedagogy and Education, 28*(2), 217–236. https://doi.org/10.1080/1475939X.2019.1572534.

Bliese, P. D., & Halverson, R. R. (1998). Group consensus and psychological well-being: A large field study. *Journal of Applied Social Psychology, 28*(7), 563–580. https://doi.org/10.1111/j.1559-1816.1998.tb01720.x.

Bradshaw, R. (2017). *Improvement in Tripod student survey ratings of secondary school instruction over three years.* Doctoral dissertation, Boston University, Boston, MA.

Brophy, J. E., & Good, T. L. (1974). *Teacher-student relationships: Causes and consequences.* Holt, Rinehart & Winston.

Burniske, J., & Meibaum, D. (2012). *The use of student perceptual data as a measure of teaching effectiveness.* Texas Comprehensive Center. https://sedl.org/txcc/resources/briefs/number_8/index.php. Accessed 5 August 2020.

Calarco, J. M. (2011). "I need help!" Social class and children's help-seeking in elementary school. *American Sociological Review, 76*(6), 862–882. https://doi.org/10.1177/0003122411427177.

Chan, D. (1998). Functional relations among constructs in the same content domain at different levels of analysis: A typology of composition models. *Journal of Applied Psychology, 83*(2), 234. https://doi.org/10.1037/0021-9010.83.2.234.

Christle, C. A., Jolivette, K., & Nelson, C. M. (2007). School characteristics related to high school dropout rates. *Remedial And Special Education, 28*(6), 325–339. https://doi.org/10.1177/07419325070280060201.

Cohen, J., & Goldhaber, D. (2016). Building a more complete understanding of teacher evaluation using classroom observations. *Educational Researcher, 45*(6), 378–387. https://doi.org/10.3102/0013189X16659442.

Connor, C. M., Morrison, F. J., Fishman, B. J., Ponitz, C. C., Glasney, S., Underwood, P. S., Piasta, S. B., Crowe, E. C., & Schatschneider, C. (2009). The ISI classroom observation system: Examining the literacy instruction provided to individual students. *Educational Researcher,* 38(2), 85–99.

Croninger, R. G., & Valli, L. (2009). "Where is the action?" Challenges to studying the teaching of reading in elementary classrooms. *Educational Researcher, 38*(2), 100–108. https://doi.org/10.3102/0013189X09333206.

Crosnoe, R. (2005). Double disadvantage or signs of resilience? The elementary school contexts of children from Mexican immigrant families. *American Educational Research Journal, 42*(2), 269–303. https://doi.org/10.3102/00028312042002269.

de Boer, A., Pijl, S. J., Post, W., & Minnaert, A. (2013). Peer acceptance and friendships of students with disabilities in general education: The role of child, peer, and classroom variables. *Social Development, 22*(4), 831–844. https://doi.org/10.1111/j.1467-9507.2012.00670.x.

Dee, T. S. (2005). A teacher like me: Does race, ethnicity, or gender matter? *American Economic Review, 95*(2), 158–165. https://doi.org/10.1257/000282805774670446.

Den Brok, P., Brekelmans, M., & Wubbels, T. (2006). Multilevel issues in research using students' perceptions of learning environments: The case of the questionnaire on teacher interaction. *Learning Environments Research, 9*(3), 199. https://doi.org/10.1007/s10984-006-9013-9.

Douglas, K. (2009). Sharpening our focus in measuring classroom instruction. *Educational Researcher, 38*(7), 518–521. https://doi.org/10.3102/0013189x09350881.

Downer, J. T., Stuhlman, M., Schweig, J., Martínez, J. F., & Ruzek, E. (2015). Measuring effective teacher-student interactions from a student perspective: A multi-level analysis. *The Journal of Early Adolescence, 35*(5–6), 722–758. https://doi.org/10.1177/0272431614564059.

Durlak, J. A., Weissberg, R. P., Dymnicki, A. B., Taylor, R. D., & Schellinger, K. (2011). The impact of enhancing students' social and emotional learning: A meta-analysis of school-based universal interventions. *Child Development, 82,* 405–432. https://doi.org/10.1111/j.1467-8624.2010.01564.x.

Echterhoff, G., Higgins, E. T., & Levine, J. M. (2009). Shared reality: Experiencing commonality with others' inner states about the world. *Perspectives on Psychological Science, 4*(5), 496–521. https://doi.org/10.1111/j.1745-6924.2009.01161.x.

Fauth, B., Decristan, J., Rieser, S., Klieme, E., & Büttner, G. (2014). Student ratings of teaching quality in primary school: Dimensions and prediction of student outcomes. *Learning and Instruction, 29,* 1–9. https://doi.org/10.1016/j.learninstruc.2013.07.001.

Feldlaufer, H., Midgley, C., & Eccles, J. S. (1988). Student, teacher, and observer perceptions of the classroom environment before and after the transition to junior high school. *The Journal of Early Adolescence, 8*(2), 133–156. https://doi.org/10.1177/0272431688082003.

Ferguson, R. F. (2012). Can student surveys measure teaching quality? *Phi Delta Kappan, 94*(3), 24–28. https://doi.org/10.1177/003172171209400306.

Flores, A. (2007). Examining disparities in mathematics education: Achievement gap or opportunity gap? *The High School Journal, 91*(1), 29–42. https://doi.org/10.1353/hsj.2007.0022.

Follman, J. (1992). Secondary school students' ratings of teacher effectiveness. *The High School Journal, 75*(3), 168–178.

Fraser, B. J. (1998). Classroom environment instruments: Development, validity and applications. *Learning Environments Research, 1*(1), 7–34.

Fraser, B. J. (2002). Learning environments research: Yesterday, today and tomorrow. *Studies in educational learning environments: An international perspective* (pp. 1–25). World Scientific.

Fraser, B. J., & McRobbie, C. J. (1995). Science laboratory classroom environments at schools and universities: A cross-national study. *Educational Research and Evaluation, 1,* 289–317. https://doi.org/10.1080/1380361950010401.

Freeman, B., & Crawford, L. (2008). Creating a middle school mathematics curriculum for English-language learners. *Remedial and Special Education, 29*(1), 9–19. https://doi.org/10.1177/0741932507309717.

Gamoran, A., & Weinstein, M. (1998). Differentiation and opportunity in restructured schools. *American Journal of Education, 106*(3), 385–415. https://doi.org/10.1086/444189.

Garza, R. (2009). Latino and white high school students' perceptions of caring behaviors: Are we culturally responsive to our students? *Urban Education, 44*(3), 297–321. https://doi.org/10.1177/0042085908318714.

Gehlbach, H. (2015). Seven survey sins. *The Journal of Early Adolescence, 35*(5-6), 883–897.

Gehlbach, H., Brinkworth, M. E., King, A. M., Hsu, L. M., McIntyre, J., & Rogers, T. (2016). Creating birds of similar feathers: Leveraging similarity to improve teacher–student relationships and academic achievement. *Journal of Educational Psychology, 108*(3), 342. https://doi.org/10.1037/edu0000042.

Gershenson, S., Holt, S. B., & Papageorge, N. W. (2016). Who believes in me? The effect of student–teacher demographic match on teacher expectations. *Economics of Education Review, 52,* 209–224. https://doi.org/10.1016/j.econedurev.2016.03.002.

Glick, W. H. (1985). Conceptualizing and measuring organizational and psychological climate: Pitfalls in multilevel research. *Academy of Management Review, 10*(3), 601–616. https://doi.org/10.2307/258140.

Gottfredson, G. D., Gottfredson, D. C., Payne, A. A., & Gottfredson, N. C. (2005). School climate predictors of school disorder: Results from a national study of delinquency prevention in schools. *Journal of Research in Crime and Delinquency, 42*(4), 412–444. https://doi.org/10.1177/0022427804271931.

Griffith, J. (2000). School climate as group evaluation and group consensus: Student and parent perceptions of the elementary school environment. *The Elementary School Journal, 101*(1), 35–61. https://doi.org/10.1086/499658.

Guion, R. M. (1973). A note on organizational climate. *Organizational Behavior and Human Performance, 9*(1), 120–125.

Hamilton, L. S., Doss, C. J., & Steiner, E. D. (2019). *Teacher and principal perspectives on social and emotional learning in America's schools: Findings from the American educator panels.* RAND Corporation.

Heine, S. J., Lehman, D. R., Peng, K., & Greenholtz, J. (2002). What's wrong with cross-cultural comparisons of subjective Likert scales? The reference-group effect. *Journal of Personality and Social Psychology, 82*(6), 903. https://doi.org/10.1037/0022-3514.82.6.903.

Holahan, C., & Batey, B. (2019). *Measuring school climate and social and emotional learning and development: A navigation guide for states and districts.* Council of Chief State School Officers. https://ccsso.org/sites/default/files/2019-03/CCSSO-EdCounsel%20SE%20and%20School%20Climate%20measurement.pdf. Accessed 5 August 2020.

Hough, H., Kalogrides, D., & Loeb, S. (2017). *Using surveys of students' social-emotional learning and school climate for accountability and continuous improvement.* Policy Analysis for California

Education. https://edpolicyinca.org/sites/default/files/SEL-CC_report.pdf. Accessed 5 August 2020.

Hoy, W. K. (1990). Organizational climate and culture: A conceptual analysis of the school workplace. *Journal of Educational and Psychological Consultation, 1*(2), 149–168. https://doi.org/10.1207/s1532768xjepc0102_4.

Kane, T. J., & Staiger, D. O. (2012). *Gathering feedback for teaching: Combining high-quality observations with student surveys and achievement gains.* Bill & Melinda Gates Foundation. http://k12education.gatesfoundation.org/resource/gathering-feedback-on-teaching-combining-high-quality-observations-with-student-surveys-and-achievement-gains-3/. Accessed 5 August 2020.

Klieme, E., Pauli, C., & Reusser, K. (2009). The Pythagoras study: Investigating effects of teaching and learning in Swiss and German mathematics classrooms. *The power of video studies in investigating teaching and learning in the classroom* (pp. 137–160). Waxmann.

Konold, T., & Cornell, D. (2015). Multilevel multitrait–multimethod latent analysis of structurally different and interchangeable raters of school climate. *Psychological Assessment, 27*(3), 1097. https://doi.org/10.1037/pas0000098.

Kostyo, S., Cardichon, J., & Darling-Hammond, L. (2018). *Making ESSA's equity promise real: State strategies to close the opportunity gap.* Learning Policy Institute. https://learningpolicyinstitute.org/product/essa-equity-promise-report. Accessed 5 August 2020.

Kyriakides, L. (2005). Drawing from teacher effectiveness research and research into teacher interpersonal behaviour to establish a teacher evaluation system: A study on the use of student ratings to evaluate teacher behaviour. *The Journal of Classroom Interaction*, 44–66.

Lacoe, J. R. (2015). Unequally safe: The race gap in school safety. *Youth Violence and Juvenile Justice, 13*(2), 143–168. https://doi.org/10.1177/1541204014532659.

Lam, A. C., Ruzek, E. A., Schenke, K., Conley, A. M., & Karabenick, S. A. (2015). Student perceptions of classroom achievement goal structure: Is it appropriate to aggregate? *Journal of Educational Psychology, 107*(4), 1102.

Lareau, A. (2002). Invisible inequality: Social class and childrearing in black families and white families. *American Sociological Review*, 747–776.

Lareau, A., & Horvat, E. M. (1999). Moments of social inclusion and exclusion race, class, and cultural capital in family-school relationships. *Sociology of Education*, 37–53. https://doi.org/10.2307/3088916.

Lazarides, R., & Watt, H. M. (2015). Girls' and boys' perceived mathematics teacher beliefs, classroom learning environments and mathematical career intentions. *Contemporary Educational Psychology, 41,* 51–61. https://doi.org/10.1016/j.cedpsych.2014.11.005.

LeClair, C., Doll, B., Osborn, A., & Jones, K. (2009). English language learners' and non–English language learners' perceptions of the classroom environment. *Psychology in the Schools, 46*(6), 568–577. https://doi.org/10.1002/pits.20398.

Levy, J., & Wubbels, T. (1992). Student and teacher characteristics and perceptions of teacher communication style. *The Journal of Classroom Interaction*, 23–29.

Levy, J., Wubbels, T., den Brok, P., & Brekelmans, M. (2003). Students' perceptions of interpersonal aspects of the learning environment. *Learning Environments Research, 6*(1), 5–36.

Lincoln, Y. S. (1995). In search of students' voices. *Theory Into Practice, 34*(2), 88–93. https://doi.org/10.1080/00405849509543664.

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.

Lipowski, F., Rakoczy, K., Drollinger-Vetter, B., Klieme, E., Reusser, K., & Pauli, C. (2009). Quality of geometry instructions and its short-term impact on students' understanding of the Pythagorean theorem. *Learning and Instruction, 19*(1), 527–537.

Little, O., Goe, L., & Bell, C. (2009). *A practical guide to evaluating teacher effectiveness.* National Comprehensive Center for Teacher Quality. https://files.eric.ed.gov/fulltext/ED543776.pdf. Accessed 5 August 2020.

Lockheed, M. (1976). Some determinants and consequences of teacher expectations concerning pupil performance. In *Beginning teacher evaluation study: Phase II*. ETS.

Lüdtke, O., Robitzsch, A., Trautwein, U., & Kunter, M. (2009). Assessing the impact of learning environments: How to use student ratings of classroom or school characteristics in multilevel modeling. *Contemporary Educational Psychology, 34*(2), 120–131. https://doi.org/10.1016/j.cedpsych.2008.12.001.

Lüdtke, O., Trautwein, U., Kunter, M., & Baumert, J. (2006). Reliability and agreement of student ratings of the classroom environment: A reanalysis of TIMSS data. *Learning Environments Research, 9*(3), 215–230. https://doi.org/10.1007/s10984-006-9014-8.

Ma, X., & Willms, J. D. (2004). School disciplinary climate: Characteristics and effects on eighth grade achievement. *Alberta Journal of Educational Research, 50*(2), 169–188.

Maehr, M. L., & Midgley, C. (1991). Enhancing student motivation: A schoolwide approach. *Educational Psychologist, 26*(3–4), 399–427. https://doi.org/10.1207/s15326985ep2603&4_9.

Marsh, H. W. (2007). Students' evaluations of university teaching: Dimensionality, reliability, validity, potential biases and usefulness. In *The scholarship of teaching and learning in higher education: An evidence-based perspective* (pp. 319–383). Springer.

Marsh, H. W., Lüdtke, O., Nagengast, B., Trautwein, U., Morin, A. J., Abduljabbar, A. S., et al. (2012). Classroom climate and contextual effects: Conceptual and methodological issues in the evaluation of group-level effects. *Educational Psychologist, 47*(2), 106–124. https://doi.org/10.1080/00461520.2012.670488.

Martínez, J. F. (2012). Consequences of omitting the classroom in multilevel models of schooling: An illustration using opportunity to learn and reading achievement. *School Effectiveness and School Improvement, 23*(3), 305–326. https://doi.org/10.1080/09243453.2012.678864.

Maulana, R., Helms-Lorenz, M., & van de Grift, W. (2014). Development and evaluation of a questionnaire measuring pre-service teachers' teaching behaviour: A Rasch modelling approach. *School Effectiveness and School Improvement, 26*(2), 169–194. https://doi.org/10.1080/09243453.2014.939198.

Mitra, D. (2007). Student voice in school reform: From listening to leadership. In D. Thiessen & A. Cook-Sather (Eds.), *International handbook of student experience in elementary and secondary school*. Springer.

Muijs, D., Kyriakides, L., van der Werf, G., Creemers, B., Timperley, H., & Earl, L. (2014). State of the art—Teacher effectiveness and professional learning. *School Effectiveness and School Improvement, 25*(2), 231–256. https://doi.org/10.1080/09243453.2014.885451.

Oakes, J. (1990). *Multiplying inequalities: The effects of race, social class, and tracking on opportunities to learn mathematics and science*. RAND Corporation.

Ostroff, C., Kinicki, A. J., & Tamkins, M. M. (2003). Organizational climate and culture. *Comprehensive Handbook of Psychology, 12,* 365–402. https://doi.org/10.1002/0471264385.wei1222.

Panorama Education. (2015). Validity brief: Panorama student survey. *Panorama Education*. https://go.panoramaed.com/hubfs/Panorama_January2019%20/Docs/validity-brief.pdf. Accessed 5 August 2020.

Perera, R. M., & Schweig, J. D. (2019). *The role of student- and classroom-based factors associated with classroom racial climate gaps*. Presented at the annual meeting of the American Educational Research Association, Toronto, Canada.

Pianta, R. C., & Hamre, B. K. (2009). Conceptualization, measurement, and improvement of classroom processes: Standardized observation can leverage capacity. *Educational Researcher, 38*(2), 109–119. https://doi.org/10.3102/0013189X09332374.

Popham, W. J. (2013). *Evaluating America's teachers: Mission possible?* Corwin Press.

Raudenbush, S. W., & Jean, M. (2014). To what extent do student perceptions of classroom quality predict teacher value added? In T. J. Kane, K. A. Kerr, & R. C. Pianta (Eds.), *Designing teacher evaluation systems* (pp. 170–201). Jossey Bass.

Reinholz, D. L., & Shah, N. (2018). Equity analytics: A methodological approach for quantifying participation patterns in mathematics classroom discourse. *Journal for Research in Mathematics Education, 49*(2), 140–177.

Roberts, K. H., Hulin, C. L., & Rousseau, D. M. (1978). *Developing an interdisciplinary science of organizations.* Jossey-Bass.

Rothstein, J., & Mathis, W. J. (2013). *Review of "Have we identified effective teachers?" and "A composite estimator of effective teaching: Culminating findings from the measures of effective teaching project".* National Education Policy Center. https://nepc.colorado.edu/sites/default/files/ttr-final-met-rothstein.pdf. 5 August 2020.

Ryan, R. M., & Grolnick, W. S. (1986). Origins and pawns in the classroom: Self-report and projective assessments of individual differences in children's perceptions. *Journal of Personality and Social Psychology, 50*(3), 550. https://doi.org/10.1037/0022-3514.50.3.550.

Schenke, K., Ruzek, E., Lam, A. C., Karabenick, S. A., & Eccles, J. S. (2018). To the means and beyond: Understanding variation in students' perceptions of teacher emotional support. *Learning and Instruction, 55,* 13–21. https://doi.org/10.1016/j.learninstruc.2018.02.003.

Schweig, J. D. (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.

Schweig, J. D., Kaufman, J. H., & Opfer, V. D. (2020). Day by day: Investigating variation in elementary mathematics instruction that supports the common core. *Educational Researcher, 49*(3), 176–187. https://doi.org/10.3102/0013189X20909812.

Schweig, J. D., Martinez, J. F., & Langi, M. (2017). *Beyond means: Investigating classroom learning environments through consensus in student surveys.* Presented at the biennial EARLI Conference for Research on Learning and Instruction, Tampere, Finland.

Schweig, J. D., Martínez, J. F., & Schnittka, J. (2020). Making sense of consensus: Exploring how classroom climate surveys can support instructional improvement efforts in science. Manuscript under review.

Seidel, T. (2006). The role of student characteristics in studying micro teaching–learning environments. *Learning Environments Research, 9*(3), 253–271.

Shindler, J., Jones, A., Williams, A. D., Taylor, C., & Cardenas, H. (2016). The school climate-student achievement connection: If we want achievement gains, we need to begin by improving the climate. *Journal of School Administration Research and Development, 1*(1), 9–16.

Simpson, A. W., & Erickson, M. T. (1983). Teachers' verbal and nonverbal communication patterns as a function of teacher race, student gender, and student race. *American Educational Research Journal, 20*(2), 183–198. https://doi.org/10.2307/1162593.

Sirotnik, K. A. (1980). Psychometric implications of the unit-of-analysis problem (with examples from the measurement of organizational climate). *Journal of Educational Measurement, 17*(4), 245–282. https://doi.org/10.1111/j.1745-3984.1980.tb00831.x.

Stapleton, L. M., Yang, J. S., & Hancock, G. R. (2016). Construct meaning in multilevel settings. *Journal of Educational and Behavioral Statistics, 41*(5), 481–520. https://doi.org/10.3102/1076998616646200.

Stipek, D. J., Givvin, K. B., Salmon, J. M., & MacGyvers, V. L. (2001). Teachers' beliefs and practices related to mathematics instruction. *Teaching and Teacher Education, 17*(2), 213–226. https://doi.org/10.1016/S0742-051X(00)00052-4.

Teh, G. P., & Fraser, B. J. (1994). An evaluation of computer-assisted learning in terms of achievement, attitudes and classroom environment. *Evaluation & Research in Education, 8*(3), 147–159. https://doi.org/10.1080/09500799409533363.

Urdan, T., & Schoenfelder, E. (2006). Classroom effects on student motivation: Goal structures, social relationships, and competence beliefs. *Journal of School Psychology, 44*(5), 331–349. https://doi.org/10.1016/j.jsp.2006.04.003.

van der Lans, R. M., van de Grift, W. J., & van Veen, K. (2015). Developing a teacher evaluation instrument to provide formative feedback using student ratings of teaching acts. *Educational Measurement: Issues and Practice, 34*(3), 18–27. https://doi.org/10.1111/emip.12078.

Voight, A., Hanson, T., O'Malley, M., & Adekanye, L. (2015). The racial school climate gap: Within-school disparities in students' experiences of safety, support, and connectedness. *American Journal of Community Psychology, 56*(3–4), 252–267. https://doi.org/10.1007/s10464-015-9751-x.

Vriesema, C. C., & Gehlbach, H. (2019). *Assessing survey satisficing: The impact of unmotivated questionnaire respondents on data quality*. Policy Analysis for California Education. https://files.eric.ed.gov/fulltext/ED600463.pdf. Accessed 5 August 2020.

Wagner, W., Göllner, R., Helmke, A., Trautwein, U., & Lüdtke, O. (2013). Construct validity of student perceptions of instructional quality is high, but not perfect: Dimensionality and generalizability of domain-independent assessments. *Learning and Instruction, 28,* 1–11. https://doi.org/10.1016/j.learninstruc.2013.03.003.

Wallace, T. L., Kelcey, B., & Ruzek, E. (2016). What can student perception surveys tell us about teaching? Empirically testing the underlying structure of the tripod student perception survey. *American Educational Research Journal, 53*(6), 1834–1868. https://doi.org/10.3102/0002831216671864.

Watkins, A. M., & Melde, C. (2009). Immigrants, assimilation, and perceived school disorder: An examination of the "other" ethnicities. *Journal of Criminal Justice, 37*(6), 627–635. https://doi.org/10.1016/j.jcrimjus.2009.09.011.

West, M. R., Buckley, K., Krachman, S. B., & Bookman, N. (2018). Development and implementation of student social-emotional surveys in the CORE districts. *Journal of Applied Developmental Psychology, 55,* 119–129. https://doi.org/10.1016/j.appdev.2017.06.001.

Wubbels, T., & Brekelmans, M. (2005). Two decades of research on teacher–student relationships in class. *International Journal of Educational Research, 43*(1–2), 6–24. https://doi.org/10.1016/j.ijer.2006.03.003.

Jonathan D. Schweig is a Social Scientist at the RAND Corporation (USA). His research focuses on the measurement of teaching quality, and features the learning environment, the development and implementation of teacher evaluation policies, and student social and emotional learning. His recent publications focus on the implications which methodological choices can have on inferences about instructional practice, classroom climate, and program effectiveness. Jonathan earned an M.A. in Curriculum and Teacher Education from Stanford University, an M.S. in Statistics from the University of California, Los Angeles, and a Ph.D. in Education from the University of California, Los Angeles (USA).

José Felipe Martínez is an Associate Professor of Social Research Methodology at UCLA's School of Education and Information Studies (USA). His research examines the application of measurement theory and methods to issues in education policy, specifically in relation to teacher, school, and program evaluation. A particular focus comprises instruments and tools for measuring classroom climate and instructional practice in mathematics and science, including teacher portfolios, student surveys, and observation protocols. His work has been supported by the National Science Foundation, and the Spencer, WT Grant, and McDonnell foundations. Dr. Martinez teaches courses on measurement, research design, and survey methodology.