Pressures of the Season: A Descriptive Look at Classroom Quality in Second and Third Grade Classrooms

Stephen B. Plank and Barbara Condliffe

B·E·R·C

Baltimore Education Research Consortium

February 2011

Without the assistance of numerous partners, this research would not have been possible. The authors extend their gratitude to the principals at the eight study schools for their support, the classroom teachers for their openness and honesty, as well as the observation team of Johns Hopkins graduate students, Morgan State affiliates, and BERC research staff members.

The Abell Foundation, The Open Society Institute-Baltimore, and the Annie E. Casey Foundation funded this study.

Table of Contents

Executive Summary i

Background 1

Methodology 3

- Research Questions 3
- Data 3
- Observation Tool - CLASS Protocol 4
- Analysis Plan 6

Findings 8

- Three Broad Domains 8
- Individual Dimensions Benchmarked to Another Urban District 9
- Differences Between Grades and Seasons 15
- Statistical Significance and Mediating Relationships 19

Discussion, Recommendations, and Conclusions 22

References 24

Appendix A: Details of Outlier Analysis 26

Appendix B: Hierarchical Linear Models for Ten Dimensions of Observational Quality 28

List of Figures

Figure 1: Typology of eight study schools by outlier status using 2006-07 and 2007-08 data for percent students enrolled in FARMs and by grade span.....2

Figure 2: Overview of CLASS domains and dimensions 4

Figure 3: The distribution of classroom quality across 23 Baltimore City classrooms.....8

Figure 4: Four dimensions of emotional support across 23 classrooms benchmarked to another urban school district 11

Figure 5: Three dimensions of classroom organization across 23 classrooms benchmarked to another urban district.....13

Figure 6: Three dimensions of Instructional Support across 23 classrooms benchmarked to another urban district.....15

Figure 7: Illustrative examples (four of ten dimensions) of change in mean dimension score between January and May 18

Pressures of the Season: A Descriptive Look at Classroom Quality in Second and Third Grade Classrooms

Executive Summary

This report presents findings from two years of classroom observation designed to help us understand the in-school experiences of students who had been first graders in eight Baltimore public schools in 2007-08. During the 2008-09 and 2009-10 school years, we conducted fieldwork to understand learning opportunities and settings for a set of students during their second and third grade years. We utilized the Classroom Assessment Scoring System (CLASS) to observe in 23 classrooms. The CLASS is an observation instrument and protocol developed by researchers at the University of Virginia (Pianta, La Paro & Hamre 2008). It has been used as a research tool in multiple observational studies of elementary school classrooms and in evaluations of specific classroom interventions (for a listing of these studies see http://www.teachstone.org/research-and-evidence/research-summary). It has also been extensively used by schools and school systems as a teacher assessment and professional development tool. For the present project, a team of observers completed 347 CLASS observation spells across 23 classrooms during the two years. In most instances, we conducted eight observation spells in each classroom over two days in January and another eight spells over two days in May.

The CLASS measures aspects of classroom quality across 10 dimensions which can be grouped into three broader domains. These domains are *Emotional Support*, *Classroom Organization*, and *Instructional Support* (further detail and descriptions provided in the coming pages). The timing of our classroom visits allows us to comment on variation by grade level (second versus third), season (January versus May), and a contingent relationship between grade and season that is likely explained by educators' reactions to the standardized assessments they and their students face each March.[1] These assessments are the relatively low-stakes Stanford Achievement Test Series (SAT10) administered to second grade students and the high-stakes Maryland School Assessment (MSA) administered to third grade students and central to the No Child Left Behind accountability framework.

Our main findings include:

- *Emotional support* in most of the 23 classrooms was in the upper half of a moderate range. In the majority of classrooms affective warmth among teachers and students, teacher sensitivity to student needs and concerns, and attention to students' interests and points of view was evident.
- *Classroom organization* was generally in a moderate to high range: In most classrooms behavioral expectations and routines were made clear by the teacher and

[1] The second grade assessment is the SAT10, administered in April in some recent years. In the school year we visited second grade classrooms (2008-09), the SAT10 was administered to first and second graders between March 9 and 13, 2009. The MSA was administered in Grades 3 through 8 between March 8 and 17 in the year we visited third grade classrooms (2009-10).

respected by students, teachers were usually proactive and successful in maintaining order, and students were generally working productively and staying on-task.

- *Instructional support* was severely lacking in most of the 23 classrooms: In most classrooms there was little evidence of higher-order concept development, high-quality dialog and feedback to students, or rich language modeling.

- Deficiencies in the dimensions of instructional and emotional support were especially pronounced during the winter in third grade classrooms.
    - We note that this is part of the extended time period leading up to students' first encounter with the high-stakes MSA.
    - We discuss what our qualitative and quantitative data suggest about the effects of this high-stakes accountability pressure on teachers and students as they create and negotiate what can broadly be called classroom quality.

Implications

If our findings reflect the realities of many other City Schools elementary classrooms, City Schools has considerable building blocks in the emotional support and classroom organization domains. This provides a strong foundation to focus on improving practices of instructional support, specifically, higher-order thinking and concept development, elaborated feedback between teachers and students about concepts and thought processes, and rich language modeling.

Test Prep must include higher-order concept development, high-quality dialog and feedback to students, and rich language modeling. Teachers and principals may need to be persuaded that students' critical thinking skills are crucial for their performance on the MSA and overall academic growth. Whole-class or small-group discussions of problem-solving strategies, thought processes, and successful approaches to items such as Brief Constructed Responses should be offered in a conceptually rich, emotionally warm, and interactive manner.

As we analyze associations between classroom quality and student achievement in the continuing phases of our larger project, we will be attentive to whether there are sufficient "middle range" levels on some of the CLASS dimensions, beyond which additional increases show little association with higher student achievement. We will also be attentive to contingent relationships whereby the benefits of higher levels in one domain (e.g., emotional support) seem to be augmented when another domain (e.g., classroom organization) meets at least some minimum threshold level.

Pressures of the Season: A Descriptive Look at Classroom Quality in Second and Third Grade Classrooms

Stephen B. Plank and Barbara Condliffe

Background

This project originated in conversations among the members of the Baltimore Education Research Consortium (BERC) Executive Committee in Summer 2007. At that time, Baltimore City Public Schools (City Schools) CEO Dr. Alonso and other members of the Executive Committee urged our research team to extend the BERC agenda to in-depth observation and data collection in schools and classrooms. One of Dr. Alonso's priorities, in particular, was to understand the learning opportunities and achievement trajectories for students who were high-achieving first graders (in an absolute, criterion-referenced sense, and also relative to others in the district, state, and nation).

A general sense among City Schools leaders was that students who had a solid foundation in basic literacy skills at the end of first grade (e.g., letter and word recognition, reading development at or above grade level) and numeracy (e.g., counting, addition and subtraction, initial exposure to geometric shapes, graphing, and problem-solving) were highly varied in their reading and mathematics achievement two or three years later, as third or fourth graders. There was a strong desire to understand the trajectories of high-achieving first graders in the subsequent years of elementary school, and to develop more complete understandings of why some students continue to thrive and other students' academic growth slows considerably. Among the obvious speculations (though far from the only working hypotheses) was that the quality of teaching practices – and consistency of teaching effectiveness across subsequent grades – varied considerably among schools and for different students. With this working hypothesis as a general guide, we made classroom observations central to our research design.

We developed a plan to study classroom settings and student achievement trajectories in eight schools, focusing on a sample of students who had scored at the 70th national percentile or above in reading and/or mathematics on a standardized test near the end of first grade (Spring 2008) as well as others who would be their classmates in second and third grade during 2008-09 and 2009-10. Details of our analytic sample of students will be included in a forthcoming report linking classroom quality to student achievement. For the present report, we direct attention to the set of eight schools we chose to study, and how the CLASS was used to assess the quality of the teaching and learning environment.

Our selection of schools was guided by several considerations. First, given the intensity of mixed-methods data collection we desired to complete as well as our limited resources, we concluded that we could commit to approximately eight schools, a number large enough to allow meaningful comparative analysis to distinguish school-level influences from classroom- and student-level influences and processes. At the same time, eight schools was a small enough number to allow interviewing principals, developing fairly deep familiarity with each school, and amassing detailed qualitative field notes to accompany quantitative information.

Secondly, our selection of schools was guided by a desire to include schools that had been positive outliers in City Schools in terms of student achievement in recent years – controlling for a wide array of student, teacher, school, and neighborhood characteristics – as well as schools that had been negative outliers. We wanted to understand the experiences of high-achieving students who were surrounded by *relatively large numbers* of classmates performing similarly and also understand the experiences of high-achieving students who were in the midst of *relatively few* classmates performing similarly.

In studying settings that served Grades 1 through 3, we needed to consider Baltimore's mix of schools serving the kindergarten through Grade 5 and kindergarten through Grade 8 span. We also needed to consider variation in the socioeconomic compositions of different schools.

Cognizant of both factors, we arrived at a plan to include one positive outlier and one negative outlier *K-5* school located *below* the local median on free- and reduced-price meal (FARM) percentage (schools with less than 73% of students enrolled in the FARM program). Additionally, we included two positive outliers and two negative outlier *K-5* schools located *above* the local FARM percentage (schools with more than 73% FARM students). Finally, we included one positive outlier and one negative outlier *K-8* school *above* the local median FARM percentage. Thus, we included six schools with relatively poor student populations (in some cases, having up to 86% FARM students) and two schools with relatively less impoverished populations, with FARM percentages of 70.9 and 63.8. Figure 1 displays the eight schools according to their status as positive or negative outliers, grade span, and socioeconomic composition.

Figure 1. Typology of eight study schools by outlier status using 2006-07 and 2007-08 data for percent students enrolled in FARMs and by grade span.

	K-5 Grade Span	K-8 Grade Span
Lower poverty (FARM percent in 2006-07 and 2007-08 below district average of 73%)	Sch A -- Positive outlier Sch A* -- Negative outlier	*No schools sampled*
Higher poverty (mean FARM percentage in SY07 and SY08 above 73%)	Sch B -- Positive outlier Sch B* -- Negative outlier Sch C -- Positive outlier Sch C* -- Negative outlier	Sch D -- Positive outlier Sch D* -- Negative outlier

Methodology

Research Questions

The main questions addressed by our analysis of the CLASS data are:

- What does instruction look like in the 23 classrooms of our study in the domains of emotional support, classroom organization, and instructional support?
- What finer-grained understandings can be gained by examining the dimensions nested within these three broad domains?
- How do the dimensions look when benchmarked to another urban district?
- What differences are observed between second and third grade classrooms, and between January and May?
    - Does accountability affect classroom instruction in a way measurable through observation?
    - What trends can be discerned?

Data

The classroom observations summarized in this report are part of a larger data collection effort completed during the 2008-09 and 2009-10 school years. Appendix A provides additional detail on the outlier analysis and selection of schools. All eight schools participated in the first year of our project (2008-09), but only five of the eight agreed to continue during the second year (2009-10). This attrition will require special attention in some of the analyses of our later reports on the larger project, but is not particularly problematic for the present report. That is, our main findings and conclusions in the present report revolve around within-school changes between January and May. Any differences between second and third grade measured levels that might be attributable to the reduced set of schools in Year 2 do not compromise the central findings and implications we feature in the remainder of this report.

In addition to 347 CLASS observation spells in 23 classrooms, we also *(1)* compiled detailed field notes after each school visit, *(2)* administered questionnaires to teachers, *(3)* conducted annual hour-long interviews with principals, and *(4)* merged classroom observations with student-level data such as Stanford Achievement Test Series (SAT10) and Maryland School Assessment (MSA) scores, attendance, school mobility, and student demographics. These additional data sources – beyond the CLASS classroom quality information – will be featured in the two forthcoming BERC reports on this larger project.

As the CLASS protocol is central to the present report, we now present more detail on its design and purposes. The CLASS is an observation instrument developed to assess classroom quality in

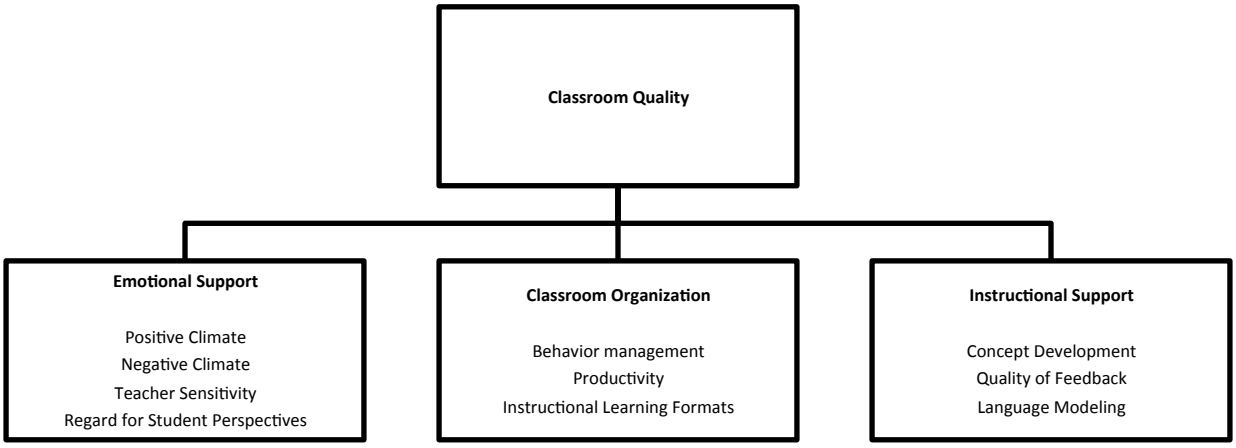
preschool through third grade classrooms. It allows us to measure classroom quality across 10 dimensions (broken into 3 broad domains) that researchers have found to be highly correlated with various measures of children's academic and social development (Mashburn et al. 2008; Hamre & Pianta 2005; Cadima et al. 2010).

By scoring each dimension separately, the CLASS elucidates specific areas of strength and weakness within each classroom. Because each dimension of the CLASS is scored on a scale of one to seven, the instrument allows us to do more than simply say that a classroom is weak or strong in a particular area. Instead we can locate a classroom along a continuum of quality. During 20 hours of training, our certified observers became calibrated to the CLASS scoring instrument and to each other in such a way that scores on a given dimension can be interpreted in an absolute, rather than a relative, sense.[2] For example, a score of 5 out of 7 on a particular dimension means that specific indicators were present in the classroom; it is not simply a reflection of the quality of this classroom in comparison to other classrooms in the study. This method of scoring allows us to make comparisons between the different classrooms we observed and between the classrooms in our study and those in other studies.

Observational Tool – CLASS Protocol

The domains of classroom quality and the dimensions measured within each domain are depicted in the following figure and discussed below. The definition of each of the ten dimensions is taken directly from the CLASS Manual (Pianta, La Paro & Hamre 2008).

Figure 2. Overview of CLASS domains and dimensions (based on Figure 1.1 in Pianta, La Paro, and Hamre, 2008).



[2] Following the training procedures and assessments of the University of Virginia developers and their Teachstone partners, every member of our classroom observation team attained a reliability rating of at least 80% on a sequence of video recordings of actual classrooms (as bench-marked against the University of Virginia and Teachstone master coders). Additionally, our team met for a series of debriefing and recalibration sessions throughout the project, to prevent "observation drift" between seasons or years.

Emotional Support

Young children's social and emotional functioning in the classroom is strongly related to their academic and social development (Pianta, Belsky, Vandergrift, Houts & Morrison 2008). The CLASS assesses four dimensions of classroom quality related to emotional support.

- **Positive Climate**: The emotional connection, respect and enjoyment demonstrated between teachers and students and among students.
- **Negative Climate**: The level of expressed negativity such as anger, hostility, or aggression exhibited by teachers and/or students in the classroom.
- **Teacher Sensitivity:** Teachers' awareness of and responsiveness to students' academic and emotional concerns.
- **Regard for Student Perspectives:** The degree to which teachers' interactions with students and classroom activities place an emphasis on students' interests, motivations and points of view.

Classroom Organization

Classroom organization is closely related to student learning. Classrooms function best and provide the most opportunities for learning when students are well-behaved, consistently have things to do, and are interested and engaged in learning tasks (Pianta, La Paro & Hamre 2008). The CLASS assesses three dimensions of classroom quality related to classroom organization.

- **Behavior Management:** How effectively teachers monitor, prevent, and redirect behavior.
- **Productivity:** How well the classroom runs with respect to routines and the degree to which teachers organize activities and directions so that maximum time can be spent in learning activities.
- **Instructional Learning Formats:** How teachers facilitate activities and provide interesting materials so that students are engaged and learning opportunities are maximized.

Instructional Support

Research suggests that *how* children are taught is critical to their academic achievement. The ways in which teachers implement the curriculum they are using and the ways in which they support cognitive and language development are critical components of a child's academic success (Pianta, La Paro & Hamre 2008). The CLASS looks at three dimensions of classroom quality related to instructional support.

- **Concept Development:** How teachers use instructional discussions and activities to promote students' higher-order thinking skills in contrast to a focus on rote instruction.
- **Quality of Feedback:** How teachers extend students' learning through their responses to students' ideas, comments, and work.
- **Language Modeling:** The extent to which teachers facilitate and encourage students' language.

Analysis

As we present descriptive findings from our CLASS observations in the next section – and relate the findings to our research questions – we proceed in four main steps.

First, we will present histograms summarizing the central tendencies of the 23 classrooms for the three broad domains (emotional support, classroom organization, and instructional support). To arrive at central tendencies, we compute the median level of each of the ten dimensions for each classroom (typically based on 16 observation spells per classroom). We then compute and plot each classroom's mean for each domain (based on the three or four computed dimension medians).[3]

The second analytic step will be to drill down to the ten dimensions in order to provide a more nuanced understanding of instruction and classroom quality in the 23 classrooms. In presenting and discussing distributions for the ten dimensions, we will make reference to both absolute levels on the CLASS's 7-point scale and comparisons with what has been found in other published studies of U.S. schools using the CLASS.

The third step will be to examine patterns by season (January versus May) as well as grade. As we present these graphs, we will recount our reasons for pursuing these analyses, which were partly motivated by *a priori* hypotheses and partly motivated by hunches or riddles that arose for us in the course of data analysis.

Finally, the fourth step will be to summarize hierarchical linear models that explicitly treat CLASS observation spells as being nested within classrooms. In developing models to predict observed levels on each of the ten dimensions for a given CLASS spell within a particular classroom, we included characteristics of the "spell" such as season, time of day, instructional format (e.g., whole class, small group, individual seat work), and content area (e.g., mathematics, English/language arts, other). We additionally explored characteristics of the classroom or school (e.g., grade level, whether the school had been selected into our study as a positive or negative outlier, and whether the school had been selected into our study as having a lower or higher FARM percentage).[4]

These *hierarchical linear modeling* (*HLM*) analyses are important for at least three reasons: First, they explicitly acknowledge the nesting of CLASS spells within particular classrooms, and thus allow us to judge whether patterns evident in the earlier descriptive graphs hold up as

[3] One of the ten dimensions – negative climate – is originally coded such that higher values are undesirable (i.e., suggest high levels of anger, hostility, aggression, and the like). In aggregating to a rating for the emotional support domain, we reverse-coded negative climate prior to calculating a mean across four dimensions. In contrast, in a histogram of negative climate in isolation, we leave intact its original coding scheme (i.e., with higher values being undesirable).

[4] Constrained by our small number of schools (and relatively small number of classrooms), it was not feasible to model statistically classrooms and schools as two distinct levels of analysis. Thus, we treat school descriptors as characteristics of classrooms in the hierarchical linear models, admittedly a slight violation of the models' logic but a step that is common in such analyses.

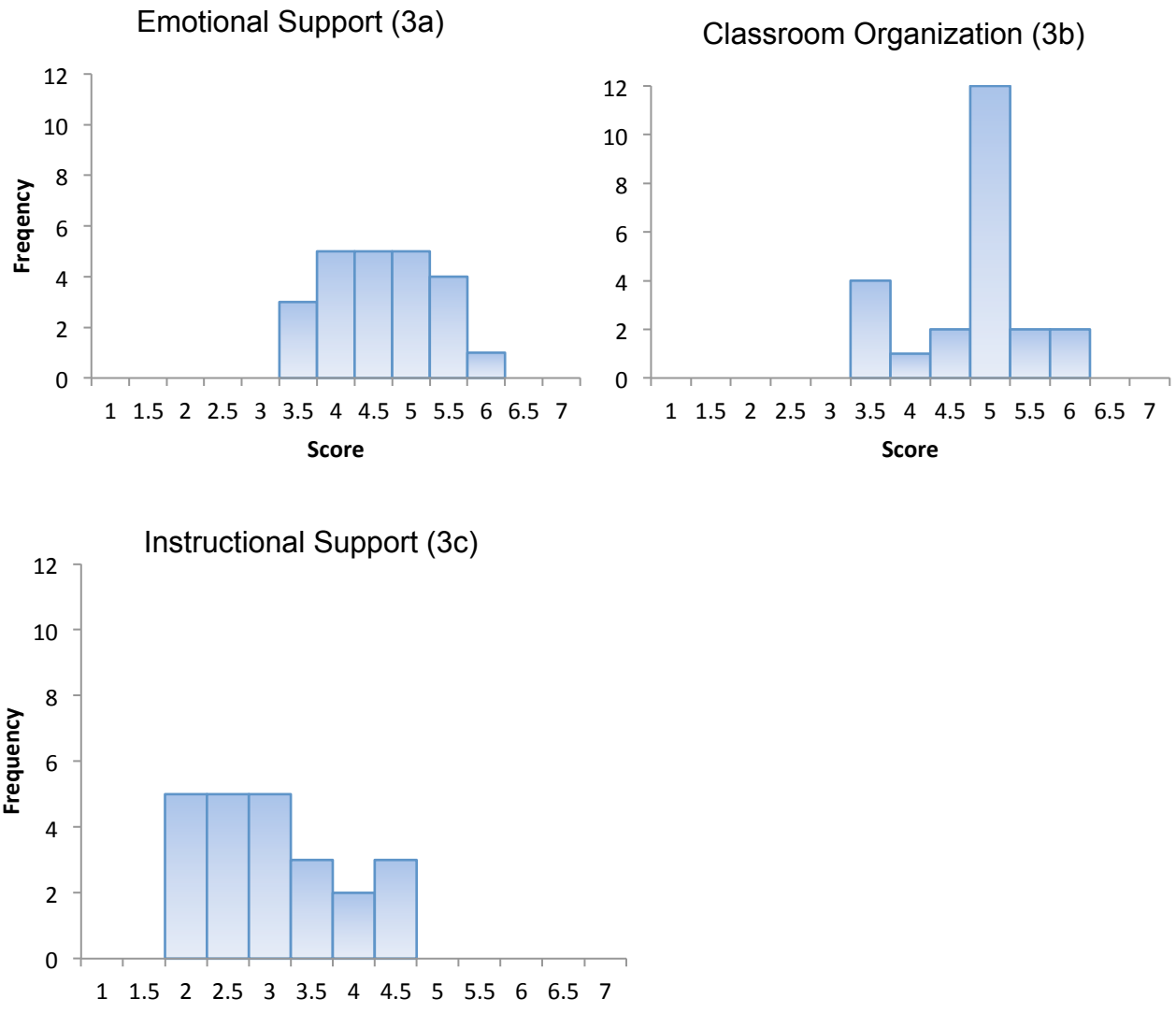
significant when subjected to appropriate statistical tests. Secondly, they provide a formal way of modeling cross-level interactions between grade (second versus third) and season (January versus May) as these relate to the ten dimensions of classroom quality – interactions that prove to be very noteworthy. Finally, the HLM analyses give us a framework for describing what seem to be important mediating relationships, in the sense of *Variable B* helping to explain the relationship between *Variables A* and *C* (as opposed to "explaining away" the relationship between *A* and *C*).

Findings

Three Broad Domains: A Look Inside 23 Classrooms

Figure 3 presents the distribution of emotional support, classroom organization, and instructional support we observed across the 23 classrooms on the CLASS's 7-point scale. Consistent with our observation team's formal training and reliability exercises, we interpret scores of 1 and 2 as being in a "low" zone (in an absolute sense). We interpret scores of 3, 4, and 5 as being in a "moderate" zone. Scores of 6 and 7 are in a "high" zone.[5]

Figure 3. The Distribution of Classroom Quality Across 23 Baltimore City Classrooms



[5] Given our aggregation to classroom averages, we will actually draw the cut-point between "low" and "moderate" at 2.5 for discussion purposes. We will draw the cut-point between "moderate" and "high" at 5.5.

As graphed in panel 3a, among the 23 classrooms the lowest aggregated rating for emotional support is 3.38 (derived by aggregating multiple observations of a particular classroom). The highest is 5.75. Twenty-two of the 23 classrooms fall within the moderate zone on emotional support, tending to be in the upper half of that zone rather than the lower. Overall then, we find that most classrooms demonstrate fairly high levels of emotional connection, teacher sensitivity to student needs and concerns, and attention to students' interests and points of view. One can imagine, and some would urge, that classrooms should rate even higher on this 7-point scale, but overall the observed levels are not distressing or antithetical to productive teaching and learning.

For classroom organization (panel 3b), the 23 classrooms ranged from a low aggregated rating of 3.17 to a high of 5.67. The distribution is much less uniform than was observed for emotional support. Indeed there is a pronounced mode between 4.5 and 5.0, with 12 of the 23 classrooms falling into that narrow band. This high-level overview of the domain is consistent with our observation team's field notes in suggesting that in most classrooms behavioral expectations and routines were shared and respected by teachers and students, teachers were fairly proactive and successful in maintaining order, and students were productive and generally on-task.

Our assessment of the classrooms changes dramatically as we move to the instructional support domain. Panel 3c shows the lowest rated classroom with a rating of 1.67 and the highest rated at only 4.33. Ten of the 23 classrooms (43%) are in the low zone on instructional support, below 2.5 on the 7-point scale. All but four classrooms are below 4.0. These numbers reiterate what our observation team's field notes consistently implied: The opportunities for higher-order thinking, the quality of feedback and inquiry between teachers and students, and the extent to which teachers modeled and encouraged challenging and rich language use in the classroom were very limited. There is room for City Schools to improve its instructional support in the elementary grades.

Individual Dimensions Benchmarked to Another Urban District

We turn now to displays of the ten finer-grained dimensions nested within the three domains, briefly summarizing the distributions for the 23 classrooms of our study and comparing them to what has been documented in two other studies conducted in urban districts that are useful points of reference.

One of these studies, useful as a reference point, is the Social and Academic Learning Study, conducted by Sara Rimm Kaufmann (2006) as a part of a three-year quasi-experimental investigation of the Responsive Classroom Approach. Data (including CLASS ratings) were collected in 88 elementary grades classrooms (Grades 1-5) in an urban district in the northeastern United States. Rimm Kaufmann's data are a useful point of comparison for City Schools in the sense that her study site shares with Baltimore an urban, east-coast location. The schools studied by Rimm Kaufman, however, had a lower percentage of racial/ethnic minority students (53.6%) than does Baltimore and a lower percentage of students receiving free or reduced-priced meals (35.3%). Also, half of the classrooms (and teachers) in Rimm Kaufmann's study were part of an intervention to explicitly work on improving practices in the domains and dimensions measured by the CLASS. Thus, these data can be taken as realistic but challenging benchmarks for City

Schools to strive for as it travels along a path of continuous improvement in classroom quality, teaching, and learning.

Rimm Kaufmann's study includes measures of eight of the ten dimensions of interest to us. For one other dimension (regard for student perspectives), we will turn to data from the Brown, Jones, LaRusso, and Aber (2010) investigation of the 4Rs Program, another intervention study, and this one conducted in 82 classrooms (grades 3-5) in New York City. As with Rimm Kaufman's study, approximately half of the classrooms, those in 9 of 18 schools, were implementing a program to improve practices related to emotional climate, classroom organization, and instructional supports.[6]

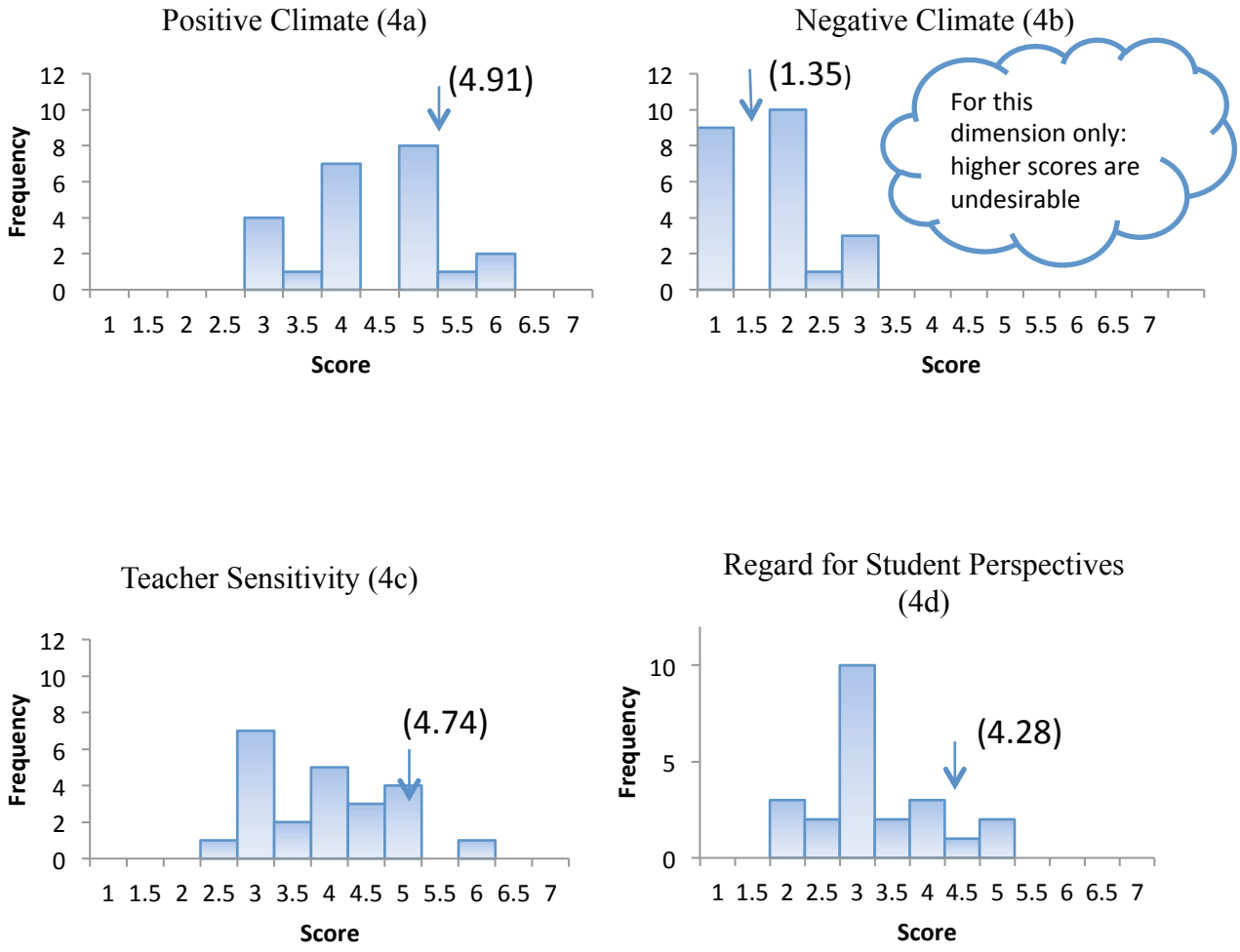
Positive Climate (part of Emotional Support). The first panel of Figure 4 (Panel 4a) shows the distribution of the 23 classrooms for positive climate. All of the classrooms were either in the moderate or high zones for their medians across our team's multiple CLASS observation spells. While the eye may be drawn to the "valley" whereby there are no observed classrooms above 4 and below 4.5, it is more important to recognize that fully 15 of the 23 classrooms had medians of 4 or 5. Thus, most of the classrooms we visited were at or slightly above the mid-point on the 7-point scale for positive climate.

In Rimm Kaufman's study, the average positive climate observed was 4.91. Twelve of the Baltimore classrooms were more than a half-point below this comparison figure. Nine were within a half-point of it, on either side. Three classrooms were more than a half-point above Rimm Kaufman's observed average. An overall summary of the Baltimore classrooms on positive climate would be: Moderate to high ranking on positive climate, certainly with room for growth or improvement if this is deemed a priority, but registering at levels that would seem to be firm foundations for teaching, learning, and future school improvement efforts.

Negative Climate (part of Emotional Support). Panel 4b shows the distribution for negative climate. Negative climate is the one dimension for which higher values are undesirable. There is no conceptual reason to expect or insist that ratings on negative climate should be perfectly negatively correlated with ratings on positive climate. That is, while in general classrooms that score highly on positive climate will score very low on negative climate (and classrooms that are mid-range on positive climate will be mid-range on negative climate), there are exceptions. Anyone who has witnessed a family that generally displays affection, love, and happiness but with periodic bursts of yelling or aggressive behavior will recognize that these two traits are not merely opposite poles of a single dimension. Analogously, our team visited a few classrooms that ranked very highly among the 23 in terms of positive climate but were far from the lowest (least problematic) on negative climate.

[6] For one of the ten dimensions of interest to us (language modeling), neither Rimm Kaufmann's data nor Brown et al.'s data include the measure, as it has been added to the CLASS protocols only in very recent years. We have not located an appropriate study to serve as a point of comparison on this dimension.

Figure 4. Four Dimensions of Emotional Support across 23 classrooms benchmarked to another urban school district.



Nineteen of the 23 classrooms had medians of either 1.0 or 2.0 for negative climate. Four classrooms were on the edge of, or just into, the moderate zone – these having medians of 2.5 and 3.0. Field notes from these four classrooms are startling; they document instances of teachers and classroom aides expressing real hostility toward children via threats, mocking, harsh criticism, and sarcasm. In this minority of classrooms, it was clear that even these isolated outbursts of aggressive behavior from the adults and sometimes from the students interfered with the development of productive interpersonal bonds and posed significant obstacles to student learning.

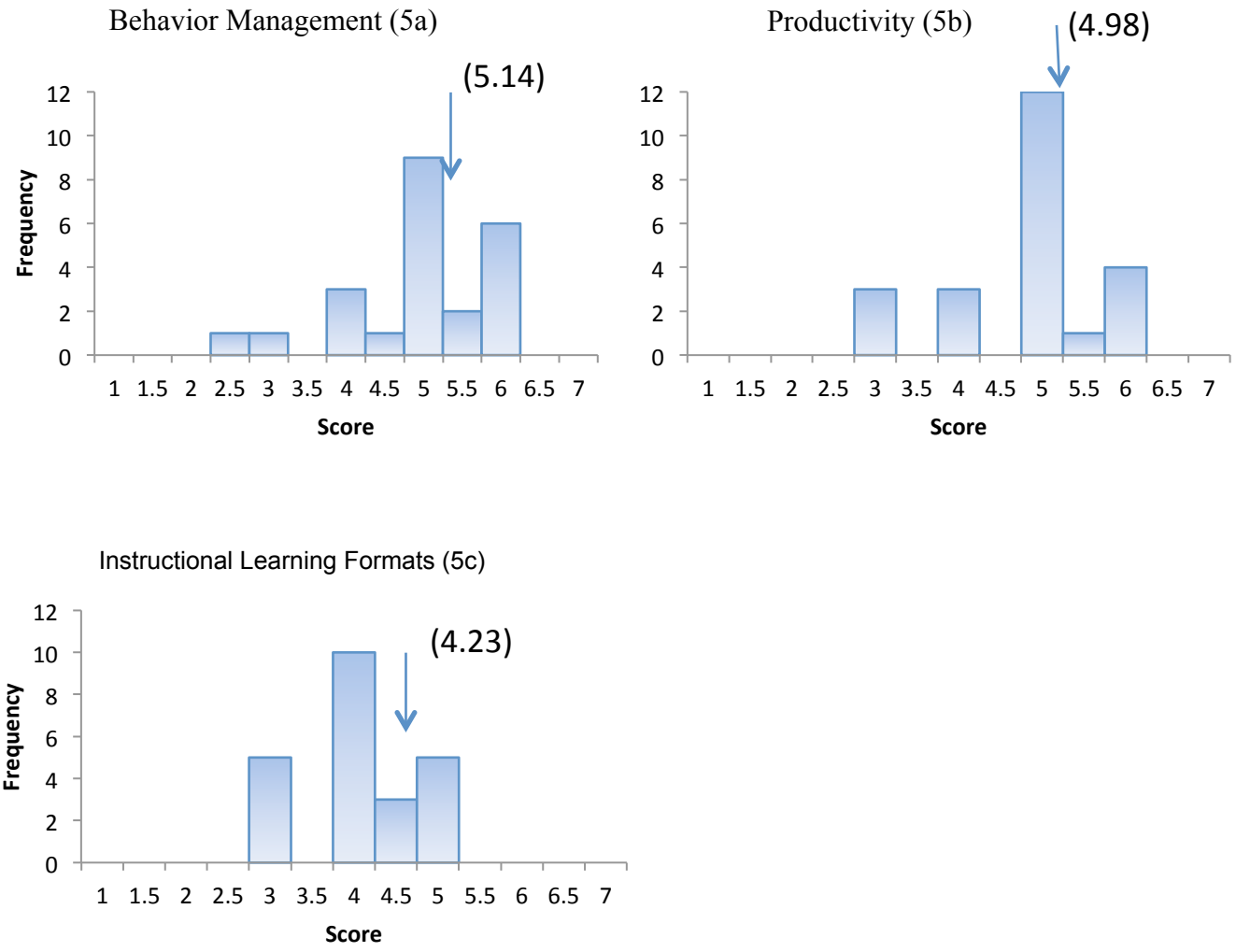
The average observed negative climate rating in Rimm Kaufman's study was 1.35. For the 14 classrooms of our study that were more than a half-point above this, and for other Baltimore classrooms with periodic moments of anger, hostility, or aggression exhibited by teachers or students, we would urge reflection and proactive strategies.

Teacher Sensitivity and *Regard for Student Perspective (parts of Emotional Support)*. Panels 4c and 4d show fairly wide distributions for teacher sensitivity and regard for student perspectives among the 23 classrooms. Relatively low ratings of classrooms in which teachers showed fairly limited awareness of and responsiveness to students' academic and emotional concerns, or placed little emphasis on students' interests, motivations, and points of view in organizing classroom activities, pulled down the classroom's overall level on the aggregate measure of emotional support. Prior research suggests that student engagement and learning are negatively affected when teacher sensitivity and regard for student perspective are at low levels (NICHD 2002; NICHD 2005). In our forthcoming report's analyses of student achievement and classroom quality, it will be important to see whether these associations are in evidence for our set of classrooms and students.

Rimm Kaufman reported an average level of teacher sensitivity of 4.74. Brown et al. reported an average regard for student perspective of 4.28. For teacher sensitivity, 15 of the 23 classrooms in our study were more than a half-point below Rimm Kaufman's reported average, and one classroom was more than a half-point above it. On regard for student perspective, 17 of the 23 classrooms were more than a half-point below Brown's (2010) mean and two were more than a half-point above it.

Behavior Management and *Productivity (parts of Classroom Organization)*. The first two panels of Figure 5 (Panels 5a and 5b) present distributions for behavior management and productivity. We discuss them as a pair because their distributions share some commonalities. Both distributions are skewed to the left, which means that there were a few classrooms with relatively low scores, located below a more concentrated group of classrooms with moderate or high scores. Fully 17 of the 23 classrooms had medians for behavior management of 5 or above. (Rimm Kaufman's reported average was 5.14.) Indeed, eight of the classrooms were in the high zone (meaning scores of 5.5 or above). Our team's field notes, as well as these scores, suggest that most of the 23 classrooms had teachers who monitored behavior effectively, and prevented or redirected off-task or resistant student behavior with relatively little disruption to other classroom activities. Further, in most classrooms, students and teachers seemed to share and utilize a set of behavioral expectations and routines. We note that strong behavioral management sometimes did, but sometimes did not, go hand-in-hand with affective warmth. In our analyses of student achievement, it will be important to investigate whether there are significant interactions between emotional support and classroom organization (including behavioral management) in predicting academic growth or achievement levels.

Figure 5. Three dimensions of classroom organization across 23 classrooms benchmarked to another urban district.



Again, seventeen of the 23 classrooms (nearly but not quite the same subset mentioned in the previous paragraph) had medians for productivity (colloquially, "organized busy-ness") of 5 or above. (Rimm Kaufman's average was 4.98.) Five of the classrooms were in the high zone (i.e., with medians of 5.5 or above). In general, our team observed classrooms in which students were usually engaged in academic activities, and teachers had established routines (including transitions between activities) to facilitate time-on-task. Of course, the details of the sorts of activities – and the nature of instructional supports flowing between teacher and students during these activities – are important and not necessarily reflected in the productivity scores; these remain to be discussed in detail, and will receive our attention in the coming paragraphs.

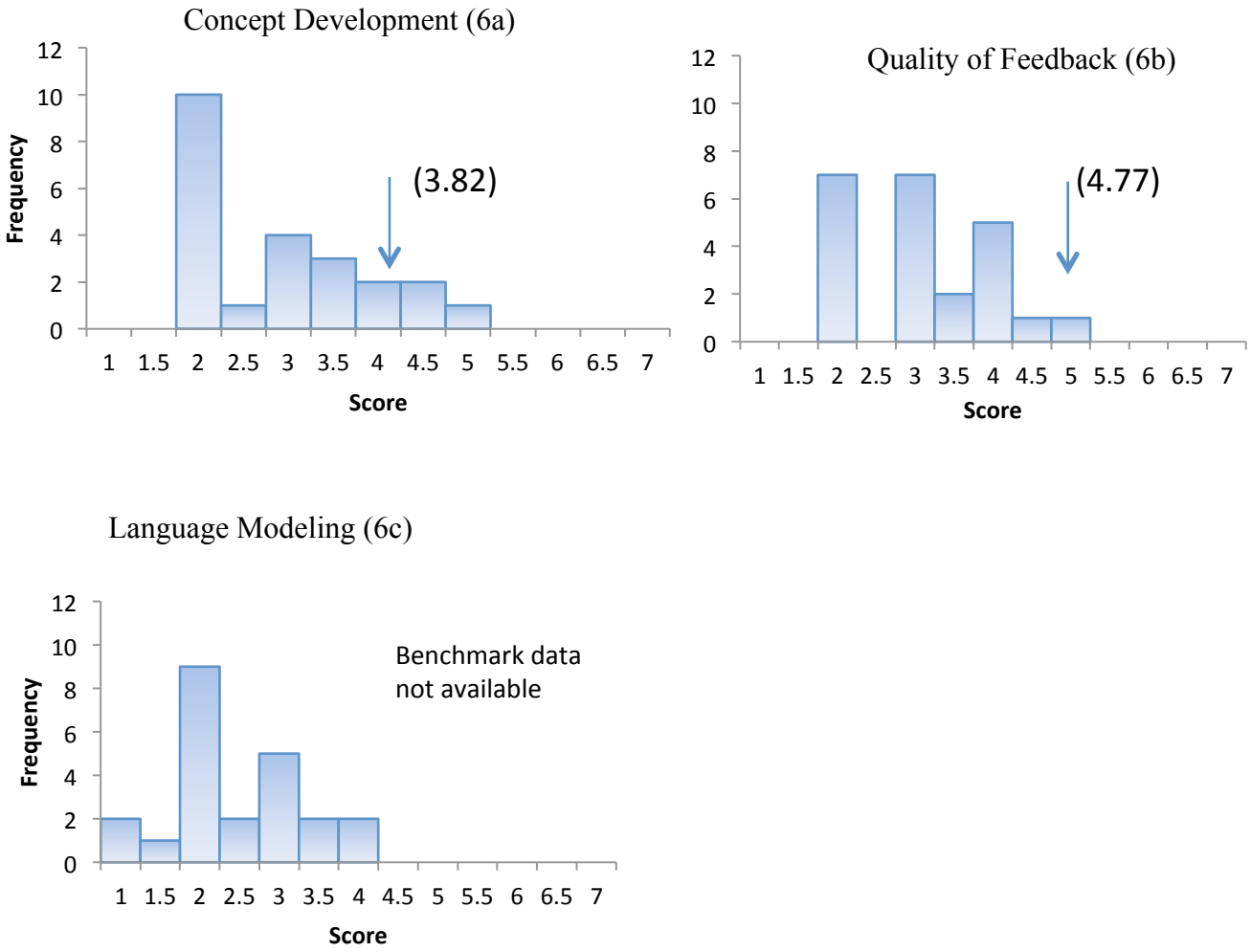
Instructional Learning Formats (part of Classroom Organization). The instructional learning format dimension (Panel 5c) reflects the extent to which teachers facilitate activities and provide materials that spark student engagement and learning opportunities. Our observation team came to think of this dimension as "things the teacher does with materials and activities to spark student engagement, and the signs of engagement itself." The 23 classrooms were all in the moderate zone (ranging between 3 and 5, specifically). Thus, we observed some positive instructional practices that utilized multiple modalities, but also saw room for improvement on this dimension.

Concept Development (part of Instructional Support). We turn now to the first of three dimensions of instructional support. The first panel of Figure 6 (Panel 6a) shows a spike or mode at 2, with 10 classrooms taking this level on the 7-point scale. The distribution is then skewed with a long tail to the right; the highest rated classroom had a median of 5. Eleven of the 23 classrooms were in the low zone on this dimension. Fifteen of the 23 were at least a half-point below Rimm Kaufman's reported average of 3.82. Our observation team saw considerably more rote instruction with a focus on mastery of basic skills of numeracy and literacy than we saw instructional practices that promote students' higher order thinking skills.

Quality of Feedback (part of Instructional Support). Panel 6b shows a distribution that is not quite as skewed as was concept development, but concentrated in the low zone and lower part of the moderate zone nonetheless. Twenty-one of the 23 classrooms had scores between 2 and 4, thus at or below the mid-point of the 7-point scale and at least a half-point below Rimm Kaufman's observed mean of 4.77. While studies conducted in other school districts attest that it is often difficult for teachers to consistently extend student learning by asking rich follow-up questions or engaging in prolonged dialogs with students, it seems clear that teachers in these 23 classrooms (and presumably others within City Schools) could use supports and encouragement regarding the sorts of responses they give to students' ideas, comments, and work (Pianta et al. 2007; Stuhlman & Pianta 2009; NICHD 2002).

Language Modeling (part of Instructional Support). Finally, Panel 6c displays the distribution for language modeling across the 23 classrooms. This dimension measures the extent to which teachers facilitate and encourage students' rich use of language. The indicators observation team members are alert for include spoken words during formal instruction as well as side conversations between teachers and students, and also printed materials posted throughout the room or incorporated into lessons. In practice, a teacher who rates high on language modeling will infuse into daily routines words that are age-appropriate and related to classroom activities, but that stretch and challenge students' current vocabularies and powers of expression.

Figure 6. Three dimensions of Instructional Support across 23 classrooms benchmarked to another urban district.



The medians for the 23 classrooms range from 1 to 4. Fourteen of the 23 classrooms were rated at 2.5 or below, thus at or below the dividing line between "low" and "moderate." Neither Rimm Kaufman nor Brown et al. measured this dimension in their research (the dimension having been added to the CLASS protocol more recently), thus we do not have a ready point of comparison from another study. However, in an absolute sense, we assert that we observed distressingly low levels of language modeling across the set of 23 classrooms.

Differences Between Second And Third Grade Classrooms and January to May

As we examine the histograms for the dimensions, it is important to remember that these charts depict classroom medians, based on aggregations of data from various times of day, with various

subject matters or instructional formats being featured, from a combination of second and third grade classrooms, and as observed in both January and May during our two-year study. The histograms are useful in providing an initial understanding of the average level of classroom quality that students in the 23 classrooms were exposed to across a substantial portion of a school year. One also wants to know, however, whether there are particular trends associated with the finer-grained contextual details (i.e., time of day, subject, instructional format, grade, and season).

Impact of Accountability

In discussing these contextual factors with educators and others knowledgeable about policy and practice in Maryland (and Baltimore, in particular), differences between the level of accountability pressure experienced by second and third grade teachers and students repeatedly arose as a theme. In March of each year, third grade students across Maryland take the MSA for the first time. Their scores on this exam have implications for their school's accountability rating and make a strong imprint on conversations and activities among principals, teachers, and students. Students in second grade take the SAT10, but it is not part of determining whether the school attains adequate yearly progress (AYP) per No Child Left Behind. Having data from January (before students take the tests) and May (after the tests) gave us the opportunity to ask whether the "shadow" of the high-stakes MSA might be seen in our data on classroom practices and quality.

We offer as an orienting logic that ***if*** *the influence of the MSA on instructional practices, classroom climate, and quality in third grade classrooms is no different from the influence of the SAT10 on practices, climate, and quality in second grade classrooms* ***then*** *January-to-May changes for any of the ten dimensions should track in parallel for the two grades (in terms of direction of change and magnitude).*

Seasonal Trends

Various hypotheses can be offered about whether emotional support, classroom organization, and instructional support will tend to improve or deteriorate between winter and spring (between January and May). On the one hand, teachers and students will have gotten to know one another – personalities, styles, emotional and learning needs, shared routines and expectations – increasingly well as the year progresses. These facts might argue for improvements on some of the ten dimensions between January and May. On the other hand, there is a well-understood fatigue, burn-out, or anticipation of summer vacation that can affect students and teachers alike, thus arguing for declining ratings on some of the ten dimensions between January and May.

We are less interested in generating – or testing – hypotheses about whether ratings on the ten dimensions (or three domains) will increase or decrease in general between January and May than we are in testing whether the direction and magnitude of change is markedly different in third grade as compared to second. If the accountability pressures of the MSA are having a marked effect on instruction in the months leading up to the assessment's administration – and an effect stronger than or different from what the SAT10 implies for second grade – this should be visible in data plots and statistical models.

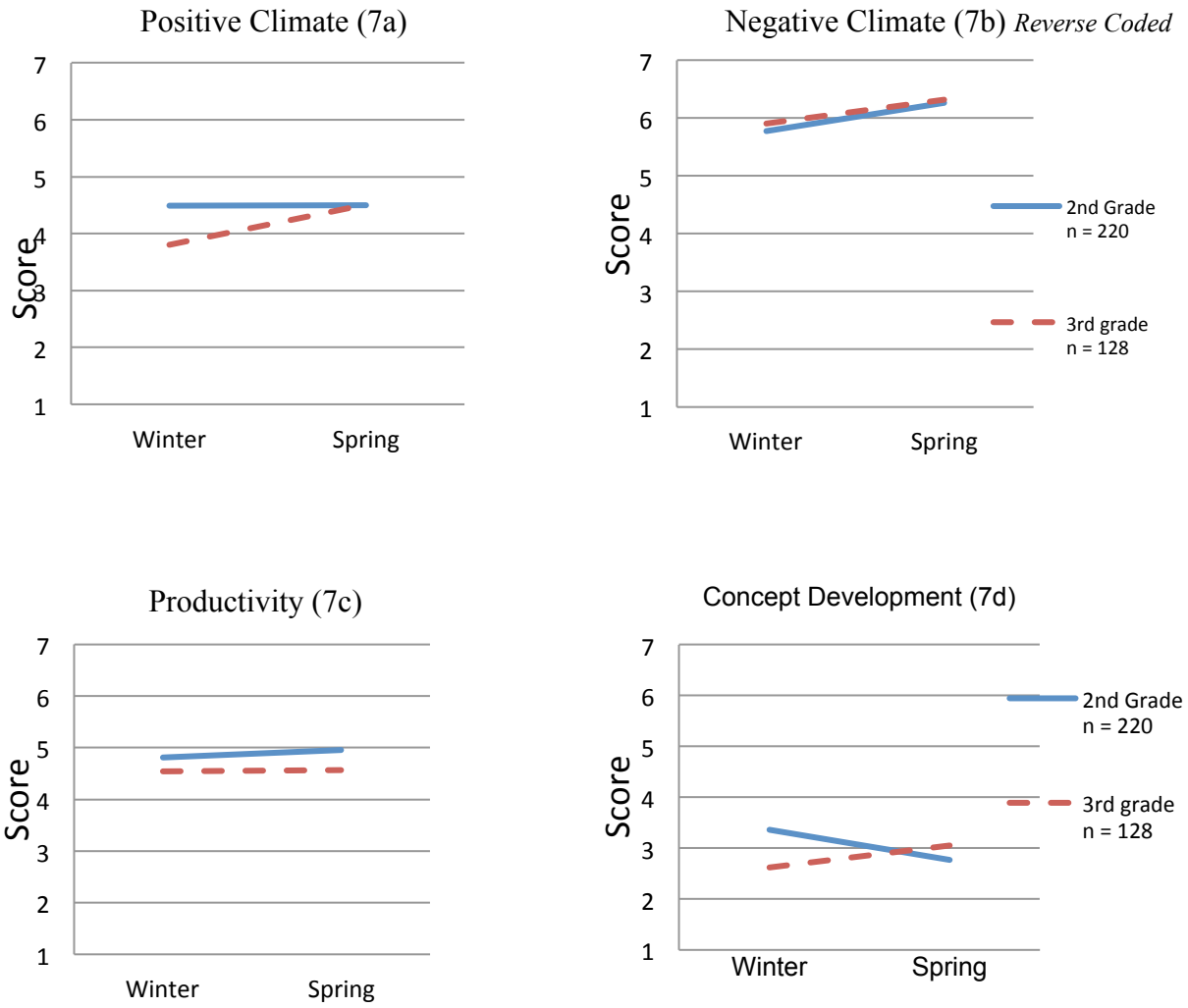
We begin by presenting descriptive charts that lump together all data points from January in second grade classrooms and then – separately – May in second grade classes, January in third grade classes, and May in third grade classes. For illustrative purposes, we present four of the ten dimensions in Figure 7, to provide important images or trends that allow us to discuss the other six dimensions as well. To confirm whether apparent trends rise to the level of statistical significance, formal models will be summarized and discussed in the next section. The panels of Figure 7, however, are useful in beginning the exploration.

Figure 7 shows January-to-May changes for positive climate, negative climate, productivity, and concept development. The results for positive climate and concept development are suggestive of a third grade story that stands in marked contrast to the second grade story. In particular, in Panel 7a, we see positive climate ratings in second grade classrooms that are quite steady between January and May – registering at about 4.5 in both seasons. In contrast, the positive climate rating increases noticeably between January and May in third grade classrooms, increasing by about three-quarters of a point from about 3.80 to 4.51. It is noteworthy that by May second and third grade levels on positive climate have come to be virtually identical.

One obvious interpretation is that with the pressures of test-preparation and the general feeling of urgency and anxiety in third-grade classrooms in the weeks and months leading up to the MSA, teachers and students are not able to generate the warmth, positive affect, and emotional connection that they might generally desire or that typifies other grades and seasons. Indeed, our observation team's field notes from the visits to third grade classrooms in January confirm that in almost all of these classrooms students were spending a good deal of their time on paper and pencil skill-based worksheets that did not require critical thinking or collaboration. In at least two schools our observers noted that instruction was purposefully organized to facilitate MSA test preparation.

Panel 7d, depicting concept development, suggests an even stronger contrast between a third grade story and a second grade story. While the second grade levels on concept development drop considerably between January and May (from 3.36 to 2.77), the third grade levels rise from 2.62 to 3.05. Our team's field notes suggest that a "winding down" of serious instruction, and a sort of spring fever, might explain the decline observed in second grade classrooms. For third grade, however, any such possible effects of spring fever or end-of-year relaxation appear to be outstripped by other forces that propel third grade classrooms toward increased focus on concept development as we move from January to May.

Figure 7. Illustrative examples (four of ten dimensions) of change in mean dimension score between January and May.



Note: While Figure 7 shows linear trends between January and May, we readily acknowledge that changes between January and May probably did not follow the straight-line patterns plotted, but instead likely followed some step-pattern with an abrupt change after the MSA administration. We do not have a steady stream of classroom observation data between January and May, but rather concentrated weeks of data collection in January and in May.

The combination of our team's field notes and the HLM results (to be presented in the next section) suggests at least two explanations for the relatively low third grade January scores and the marked improvement between January and May in third grade classrooms in many of the dimensions of classroom quality: First, third grade teachers may have been concerned with shoring up students' basic literacy and numeracy skills in advance of the MSA to such an extent that delving into higher-order problem solving, or discussing students' approaches to how they

would analyze and approach a writing prompt or mathematics story problem, might have been squeezed out of instructional activities. Second, the very nature of classroom activities related to test preparation may have precluded teacher-led activities that facilitated concept development.

Our team observed a large amount of independent seatwork when students were taking practice assessments or engaged in activities related to basic skill development. In what might be deemed a lost opportunity, however, we did not observe a lot of discussion about how one approaches a writing prompt or solves a complex mathematics question. Rather, we were much more likely to see students working silently on worksheets. We see fairly pronounced patterns akin to what is seen for positive climate or concept development for teacher sensitivity, regard for student perspective, and quality of feedback.

Panels 7b and 7c, featuring negative climate and productivity, do not seem to show divergent trends for the two grades. We include these plots in order to offer an even-handed overview. That is, descriptive plots do not reveal differential trends by grade for all of the ten dimensions. We see no pronounced grade-related differences in the January-to-May changes (direction or magnitude) for negative climate, behavioral management, and productivity. While we have some worries about the instructional climate in third grade classrooms in the weeks and months leading up to the MSA, a lack of behavioral order or a failure to have students on-task with activities are not among these worries. Finally, instructional learning formats and language modeling fall into a middle range based on visual inspection of graphs like those of Figure 7. That is, for instructional learning formats and language modeling we see some divergent seasonal trends between second and third grades, but not as pronounced as what is seen for five other dimensions. All of these initial summaries of grade-by-season patterns await more formal treatment via hierarchical linear models. We turn to these models now.

Statistical Significance and Mediating Relationships via Hierarchical Linear Models

For a more complete reporting of the HLM results, see Appendix B. The explanatory variables featured in Tables B.1 through B.10 (though dropped from some model estimations when non-significant) are season, time of day, instructional format, content area, and grade level. We also explored a school's status as a K-5 versus K-8 school and its status as a positive or negative outlier in our study design. A school's grade span was not significant in predicting classroom quality for any of the ten dimensions. Positive or negative outlier status was marginally significant for three of the ten dimensions, but not in consistent directions. That is, in two instances classrooms within the positive outlier schools were, on average, below classrooms within the negative outlier schools on a dimension. In the third instance, classrooms within the positive outlier schools were above classrooms within the negative outlier schools. The inconsistent pattern of findings raises interesting questions about schools' dynamic reactions to prior performance or close scrutiny from district and state administrators, as well as changes in staffing, to be explored in future reports.

Table 1 summarizes what the hierarchical linear models revealed regarding three key questions:

- For each of the ten dimensions, were third grade classrooms significantly different from (and, in particular, weaker than) second grade classrooms as measured by our team in January?
- Were third grade classrooms significantly different from (again, weaker than) second grade classrooms in May?
- Was the direction and/or magnitude of change between January and May significantly different in third grade classrooms, as compared with second grade classrooms?

Table 1. Summary of Statistically Significant Findings for Grade and Season (based on baseline models of Figures B.1 through B.10)

Dependent Variable	Significant difference in January? (Gr 2 vs. Gr 3)	Significant difference in May? (Gr 2 vs. Gr 3)	Significantly different change January to May? (Gr 2 vs. Gr 3)
Positive Climate	**yes ***	no	no
Negative Climate	no	no	no
Teacher Sensitivity	**yes ***	no	no
Regard for Student Perspectives	**yes ***	no	no
Behavioral Management	no	no	no
Productivity	no	no	no
Instructional Learning Formats	**yes ****	no	**yes ***
Concept Development	**yes ****	no	**yes****
Quality of Feedback	**yes ****	no	**yes ***
Language Modeling	no	no	**yes ***

* $p < 0.10$
** $p < 0.05$

In answer to the first of these questions, January differences between second and third grade classrooms significant at the $p<0.10$ level or $p<0.05$ level are found for six of the ten dimensions: positive climate, teacher sensitivity, regard for student perspective, instructional learning formats, concept development, and quality of feedback. In each case, the second grade levels were significantly higher (more desirable) than were the third grade levels. This is initial evidence that something about the mid-winter months in third grade (with MSA administration six or eight weeks in the offing) was associated with (a) less warmth, sensitivity, and regard for student perspective in the emotional domain, (b) less rich conceptual development and teacher-student feedback in the instructional domain, and (c) learning formats that were more limited in scope or less effective in sparking student engagement.

In answer to the second question, no significant May differences between second and third grade classrooms were found. The implications of this are important: After the SAT10 and MSA had been completed, second and third grade classrooms became indistinguishable from one another

across these ten dimensions, whereas that had not been true three or four months earlier. We take this as evidence that teachers and students are reacting to their environments (assessment and accountability environments, among other possibilities) in negotiating interpersonal relations, classroom routines, instructional modes, and classroom quality more generally. When environments or pressures are more similar between second and third grade classrooms (that is, in May), classroom quality is more similar. When environments or pressures distinguish second and third grade classrooms from one another (that is, in the months leading up to March), classroom quality differs between the grades.

Finally, in answer to the third question, for the three dimensions of instructional support (concept development, quality of feedback, and language modeling) and for instructional learning formats, third grade classrooms had a positive change on average between January and May while second grade classrooms had a negative change. Furthermore, these divergent January-to-May patterns rose to the level of being statistically significantly different (at the level of $p<0.05$ in one instance and $p<0.10$ in the other instances).

We note that the dimensions of emotional support that had significant January differences between second and third grade classrooms did not have significantly different January-to-May patterns of change when second and third grades are compared. These findings do not undermine our general argument – that the winter months in third grade classrooms bring a muted emotional connection and a narrowing of instructional focus and richness. The pattern of findings does, however, suggest that the phenomenon we are reporting on plays out most strongly in the instructional support domain, as well as in instructional learning formats.

The estimated hierarchical linear models also reveal what appear to be interesting and important mediation processes. That is, it seems some of the association between a particular season or grade and a measure of classroom quality can be explained via the particular instructional formats or subject areas being featured in that season or grade. To be concrete, the muted classroom quality ratings observed for third grade classrooms in January are partially accounted for by the fact that independent seatwork (very often, though not exclusively, completion of test preparation activities or practice tests) was used more often in third grade than in second, and in particular during January in those third grade classrooms. It is true in general (in our data) that classroom quality ratings (especially for the dimensions of instructional support) tend to be significantly lower during independent seatwork than during whole-class instruction or when small-group learning arrangements were employed. Similarly, the models provide some suggestion that the muted classroom quality ratings observed for third grade classrooms in January are partially accounted for by less teaching of language arts and mathematics in that grade and season and increased time dedicated to "other" subjects (a category used by our observation team for practice test-taking as well as a few other activities that fell outside the traditional academic areas). This explanation is completed by the knowledge that instructional support ratings, as well as emotional support ratings, tended to be higher when language arts were being taught than when other subjects were being taught.[7]

[7] Additional details of mediating relationships among the variables, and what can be learned by comparing the baseline models of Appendix B (column 1 of each Figure) to the final models (column 2 of each Figure), are available for inspection by interested readers, or by requesting additional information from this report's authors.

Discussion, Recommendations, and Conclusions

We offer the following points as discussion and possible implications of this descriptive report:

- The findings from these 23 classrooms (and especially if they reflect the realities of many other City Schools elementary classrooms, which we suspect they do) suggest that fairly good building blocks exist in the emotional support and classroom organization domains. Therefore, there is a strong foundation on which to develop improved practices of instructional support, higher-order thinking and concept development, feedback between teachers and students about concepts and thought processes, and language modeling.

- This is not to say that every aspect of emotional support and classroom organization was observed at high levels by our research team. For example, we witnessed multiple instances of very negative affect, yelling, hostility, and anger that seemed to seriously undermine trust and productive relationships between teachers and students, and that seemed to undermine teaching and learning more generally. For the subset of educators who allow themselves and their students to engage in bouts of negativity and aggression, explicit and proactive strategies (to modify the interpersonal behaviors of teachers and students alike) are a priority.

- We imagine that successful strategies for improving the capacity of teachers to offer rich concept development, high-quality feedback to students, and language modeling will come through a combination of recruiting and retaining people who excel in these areas in their own intellectual work, as well as developing instructional capacity among current teachers who have limited pedagogical repertoires in these areas.

- We understand the genuine dilemma teachers and principals face as they and their students approach the MSA for the first time in third grade. There is an anxiety and sense of urgency about shoring up basic literacy and numeracy skills, as well as simply having students familiar with the format and environment of MSA administration. While these realities are undeniable, we offer several suggestions that could prevent the tendencies we observed in third grade classrooms during January (that is, a heavy emphasis on students sitting for practice exams, a narrowing of the curriculum and downgrading of emphasis on higher-order thinking and concept development, and less warmth in emotional climate at a very time students might benefit from encouragement and tight bonds with their teachers and classmates). We suggest:

    - Coordinating efforts across the kindergarten to Grade 3 span so that basic literacy and numeracy are being developed early and on a continuous basis so that third grade teachers do not bear the burden of shoring up neglected basic skills – to the detriment of higher-ordering thinking skills.

    - Persuading (or confirming for) teachers that the promotion of students' critical thinking skills is crucial for their academic growth and their performance on

the MSA. Specifically, we recommend that district educators work together to develop curriculum and instructional strategies that build students' capacity to articulate their thought processes and problem-solving strategies in speaking and writing.

- Communicating to principals and teachers that preparation for the high-stakes MSA is worthwhile and to be expected, but that whole-class or small-group discussions of problem-solving strategies, thought processes, and successful approaches to items such as Brief Constructed Responses should be offered in a conceptually rich, emotionally warm, and interactive manner. Having students work independently (and silently) on practice exams or test preparation exercises many weeks before the MSA administration is probably not a good educational investment.

- Our summary of CLASS data has perhaps implied that "more is better" for each of the domains and dimensions. We recognize that such a claim is probably naïve or problematic. Teachers face difficult trade-offs and decision points. For example, it may be that teachers being responsive to students' perspectives and desires to have a voice in classroom activities is beneficial to student engagement and achievement, but taken too far (or without carefully established classroom norms and routines) regard for student perspective can interfere with time-on-task and behavioral management.

- As we analyze associations between classroom quality and student achievement (in the next stages of our larger project), we will explore whether there are sufficient "middle range" levels on some of the CLASS dimensions, beyond which additional increases show little association with higher student achievement. We will also explore contingent relationships whereby the benefits of higher levels in one domain (e.g., emotional support) seem to be augmented when another domain (e.g., classroom organization) meets at least some minimum threshold level.

- Finally, we view the portions of this paper that examine differences in classroom quality by grade and by season as contributing to broader research efforts to understand when, where, and how major policy initiatives and accountability paradigms infiltrate and affect the technical core of curriculum and instruction. Such infiltrations and effects are far from automatic given the decoupling and buffering that often characterizes schools and school systems, but they have been documented and interpreted in various times and places (e.g., Bidwell 2001; Booher-Jennings 2005; Diamond 2007; Krieg 2008). While our report is focused specifically on the contexts of Baltimore and Maryland, it is clear that most states, districts, and schools across the United States are facing similar dynamics and decision-points as the technical core of curriculum and instruction encounters shifting assessments and accountability contexts. We intend for our ongoing research to contribute to better understanding the pertinent theoretical and practical issues from both a Baltimore-based and comparative perspective.

References

Bidwell, C.E. (2001). Analyzing schools as organizations: Long-term permanence and short-term change. *Sociology of Education*, *74*, 100-114.

Booher-Jennings, J. (2005). Below the bubble: Educational triage and the Texas accountability system. *American Educational Research Journal*, *42*, 231-268.

Brown, J.L., Jones S.M., LaRusso ,M.D. & Aber J.L.(2010). Improving classroom quality: Teacher influences and experimental impacts of the 4Rs program. *Journal of Educational Psychology*, *102*, 153-167.

Cadima, J., Leal, T. & Burchinal, M. (2010). The quality of teacher-student interactions: Associations with first graders' academic and behavioral outcomes." *Journal of School Psychology*, *48*, 457-482.

Diamond, J.B. (2007). Where the rubber meets the road: Rethinking the connection between high-stakes testing policy and classroom instruction. *Sociology of Education, 80,* 285-313.

National Institute of Child Health and Human Development and Early Child Care Research Network (NICHD). (2002). The relation of global first-grade classroom environment to structural classroom features and teacher and student behaviors. *The Elementary School Journal*, *102,* 367-387.

National Institute of Child Health and Human Development Early Child Care Research Network (NICHD). (2005). A day in third grade: A large-scale study of classroom quality and teacher and student behavior. *The Elementary School Journal 105,* 305-323.

Hamre, B.K., & Pianta, R.C.(2005). Can instructional and emotional support in the first-grade classroom make a difference for children at risk of school failure? *Child Development 76,* 949-967.

Krieg, J.M. (2008). Are students left behind? The distributional effects of the No Child Left Behind act. *Education Finance and Policy, 3*, 250-281.

Mashburn, A.J.,Pianta, R.C., Hamre, B.K., Downer, J.T., Barbarin, O.A. Bryant, D., Burchinal,M., Early,D.M & Howes, C.(2008). Measures of classroom quality in prekindergarten and children's development of academic, language, and social skills." *Child Development*, *79*, 732-749.

Pianta, R.C., Belsky, J., Houts, R.& Morrison, F. (2007). Teaching: Opportunities to learn in America's elementary classrooms. *Science 315*, 1795-1796.

Pianta, R.,C. Belsky, J., Vandergrift, N., Houts, R. & Morrison, F. (2008). Classroom effects on children's achievement trajectories in elementary school. *American Educational Research Journal, 45*, 365-397.

Pianta, R.C., La Paro, K.M. & Hamre, B.K. (2008). *The Classroom Assessment Scoring System: Manual K-3.* Maryland: Paul H. Brooks Publishing Company.

Rimm-Kaufman, S.E (2006) Social and academic learning study on the responsive classroom approach. Retrieved from http://www.responsiveclassroom.org/pdf_files/sals_booklet_rc.pdf.

Stuhlman, M.W. & Pianta, R.C. (2009). Profiles of educational quality in first grade. *The Elementary School Journal 109*: 323-342.

APPENDIX A: Details of Outlier Analysis and Selection of Eight Schools

To select eight schools for our study, consistent with the logic of Figure 1 and accompanying text, we used the following procedures:

Step 1. With data from all schools serving the elementary grades in 2008-09 (and having been in operation during 2006-07 and 2007-08), we sought a single composite measure of student academic performance during the two years preceding our scheduled data collection (thus, during 2006-07 and 2007-08). To generate such a composite measure, we conducted a factor analysis (using varimax rotation) of eight school-level achievement indicators. These indicators were:

a. Percent of students proficient or advanced on Reading MSA across 3rd, 4th, and 5th grades in Spring 2007,
b. Percent proficient or advanced on Reading MSA across 3rd, 4th, and 5th grades in Spring 2008,
c. Percent proficient or advanced on Math MSA across 3rd, 4th, and 5th grades in Spring 2007,
d. Percent proficient or advanced on Math MSA across 3rd, 4th, and 5th grades in Spring 2008,
e. Average reading scale score among 2nd graders on SAT10 in Spring 2007,
f. Average reading scale score among 2nd graders on SAT10 in Spring 2008,
g. Average math scale score among 2nd graders on SAT10 in Spring 2007, and
h. Average math scale score among 2nd graders on SAT10 in Spring 2008.

The factor analysis yielded a single factor score for each school.

Step 2. This factor score was then regressed on seventeen school- and neighborhood-level characteristics:

a. Percent students at the school eligible for FARM in 2005-06,
b. Percent of students at the school identified as minority in 2005-06,
c. Log of the average 3rd grade enrollment size for 2007 and 2008,
d. Average of two dummy variables, each for whether school had an 8th grade in 2007 and 2008,
e. Rate of domestic violence incidents in 2006 (census tract),
f. Rate of juvenile arrests in 2005 (census tract)
g. Rate of juvenile violent arrests in 2005 (census tract)
h. Rate of juvenile drug arrests in 2005 (census tract)
i. Overall crime rate in 2006 (census tract)
j. Violent crime rate in 2006 (census tract)
k. Percent of houses with more than $5000 rehab investment, 2006 (census tract)
l. Percent houses vacant, 2006 (census tract)
m. Total number of housing units sold, 2005 (census tract)
n. Percent of units owner-occupied, 2006 (census tract)
o. Rate of evictions, 2005 (census tract)
p. Rate of properties under foreclosure, 2006 (census tract)
q. Total number of residential properties, 2006 (census tract)

Step 3. From this regression, studentized residuals were saved, and schools were sorted by residual. Eight schools were selected. We selected the four schools with the lowest (most negative) studentized residuals (these having values of -3.45, -1.96, -1.58, and -1.45, respectively). These are considered the negative outliers (or, colloquially, "struggling schools").

We selected four of the eight schools with the highest (most positive) studentized residuals. Specifically, we selected the schools with the first, second, fourth, and eighth highest values (3.03, 2.94, 2.03, and 1.58, respectively). These are considered the positive outliers (or, colloquially, "beacon schools"). We did not select for our study the schools in places three, five, six, or seven because we sought to pair each negative outlier in a one-to-one fashion with a positive outlier in terms of geographic region of the city, neighborhood type more generally, FARM percentage, and presence/absence of magnet or special admissions programs.[8] The schools in places three, five, six, and seven were markedly different from any of our negative outliers, and thus not optimal for inclusion in our comparative research plans.

[8] This effort at one-to-one pairing was done over and above the statistical adjustments implied by the seventeen regressors.

APPENDIX B: Hierarchical Linear Models for Ten Dimensions of Classroom Quality

Table B.1. Hierarchical Linear Models for Positive Climate

	Intercepts and Slopes as Outcomes With Season and Grade			Final Model With Additional L-1 Covariates			
Fixed Effects	Coefficient	s.e.	p-value	Fixed Effects	Coefficient	s.e.	p-value
Model for Level-1 intercept, β_{0j}				Model for Level-1 intercept, β_{0j}			
Intercept, γ_{00}	4.5 2	0.30	0. 000	Intercept, γ_{00}	4.54	0.32	0.000
Grade 3, γ_{01}	-0.75	0.42	0.091	Grade 3, γ_{01}	-0.66	0.41	0.120
Model for Spring, β_{1j}				Model for Spring, β_{1j}			
Intercept, γ_{10}	-0.01	0 .36	0.979	Intercept, γ_{10}	0.04	0.34	0.901
Grade 3, γ_{11}	0.68	0.46	0.155	Grade 3, γ_{11}	0.58	0.43	0.185
				Time			
				Mid-morning, γ_{20}			
				Afternoon, γ_{30}			
				Format			
				Small group, γ_{40}	-0.10	0. 2 2	0.653
				Independent Work, γ_{50}	-0.4 4	0.24	0.066
				Multiple Formats, γ_{60}	0.05	0.19	0.809
				Content area			
				English, γ_{70}			
				Math , γ_{80}			

Random Effects	Varianc e Component	df	χ^2	p-value
u_{0j}	1.006	21	1 43.16	0.0 0 0
u_{1j}	1.145	21	83.19	0.000
r_{ij}	1.381			

Random Effects	Variance Component	df	χ^2	p-value
u_{0j}	1.011	21	144.76	0.000
u_{1j}	1.074	21	79.19	0.000
r_{ij}	1.3 72			

Table B.2. Hierarchical Linear Models for Negative Climate (reverse coded)

	Intercepts and Slopes as Outcomes With Season and Grade			Final Model With Additional L-1 Covariates			
Fixed Effects	Coefficient	s.e.	p-value	Fixed Effects	Coefficient	s.e.	p-value
Model for Level-1 intercept, β_{0j}				Model for Level-1 intercept, β_{0j}			
Intercept, γ_{00}	5.79	0.2 2	0.000	Intercept, γ_{00}	5.79	0.22	0.000
Grade 3, γ_{01}	0.16	0.31	0.61 4	Grade 3, γ_{01}	0.16	0.31	0.614
Model for Spring, β_{1j}				Model for Spring, β_{1j}			
Intercept, γ_{10}	0.46	0.19	0.022	Intercept, γ_{10}	0.46	0.19	0.022
Grade 3, γ_{11}	-0.08	0.30	0.799	Grade 3, γ_{11}	-0.08	0.30	0.799
				Time			
				Mid-morning, γ_{20}			
				Afternoon, γ_{30}			
				Format			
				Small group, γ_{40}			
				Independent Work, γ_{50}			
				Multiple Formats, γ_{60}			
				Content area			
				English, γ_{70}			
				Math , γ_{80}			

Random Effects	Variance Component	df	χ^2	p-value
u_{0j}	0.537	21	143.77	0. 00 0
u_{1j}	0.335	21	54 .47	0.000
r_{ij}	0.746			

Random Effects	Variance Component	df	χ^2	p-value
u_{0j}	0.537	21	143.77	0.000
u_{1j}	0.335	2	54.47	0.000
r_{ij}	0.746			

Table B.3. Hierarchical Linear Models for Teacher Sensitivity

	Intercepts and Slopes as Outcomes With Season and Grade			Final Model With Additional L-1 Covariates			
Fixed Effects	Coefficient	s.e.	p-value	Fixed Effects	Coefficient	s.e.	p-value
Model for Level-1 intercept, β_{0j}				Model for Level-1 intercept, β_{0j}			
Intercept, γ_{00}	4.22	0.24	0.000	Intercept, γ_{00}	3.96	0.28	0.000
Grade 3, γ_{01}	-0.69	0.38	0.084	Grade 3, γ_{01}	-0.65	0.38	0.106
Model for Spring, β_{1j}				Model for Spring, β_{1j}			
Intercept, γ_{10}	-0.11	0.33	0.749	Intercept, γ_{10}	-0.07	0.33	0.841
Grade 3, γ_{11}	0.60	0.41	0.15 7	Grade 3, γ_{11}	0.58	0.42	0.176
				Time			
				Mid-morning, γ_{20}			
				Afternoon, γ_{30}			
				Format			
				Small group, γ_{40}			
				Independent Work, γ_{50}			
				Multiple Formats, γ_{60}			
				Content area			
				English, γ_{70}	0.30	0.15	0.039
				Math , γ_{80}	0.33	0.21	0.111

Random Effects	Variance Component	df	χ^2	p-value
u_{0j}	0.7 31	21	126.09	0.000
u_{1j}	0.922	21	80.21	0.000
r_{ij}	1.180			

Random Effects	Variance Component	df	χ^2	p-value
u_{0j}	0.724	21	125.54	0.000
u_{1j}	0. 945	21	82.24	0.000
r_{ij}	1.172			

Table B.4. Hierarchical Linear Models for Regard for Student Perspectives

	Intercepts and Slopes as Outcomes With Season and Grade			Final Model With Additional L-1 Covariates			
Fixed Effects	Coefficient	s.e.	p-value	Fixed Effects	Coefficient	s.e.	p- value
Model for Level-1 intercept, β_{0j}				Model for Level-1 intercept, β_{0j}			
Intercept, γ_{00}	3.47	0.20	0.000	Intercept, γ_{00}	3.55	0.22	0.000
Grade 3, γ_{01}	-0.55	0.29	0.077	Grade 3, γ_{01}	-0.59	0.29	0.0 56
Model for Spring, β_{1j}				Model for Spring, β_{1j}			
Intercept, γ_{10}	0.13	0.31	0.680	Intercept, γ_{10}	0.11	0.31	0.716
Grade 3, γ_{11}	0.58	0.40	0.163	Grade 3, γ_{11}	0.60	0.40	0.149
				Time			
				Mid-morning, γ_{20}	0.06	0.19	0.739
				Afternoon, γ_{30}	-0.29	0.16	0.075
				Format			
				Small group, γ_{40}			
				Independent Work, γ_{50}			
				Multiple Formats, γ_{60}			
				Content area			
				English, γ_{70}			
				Math , γ_{80}			

Random Effects	Variance Component	df	χ	p-value
u_{0j}	0.374	21	63.43	0.0 0 0
u_{1j}	0.754	21	60.51	0.000
r_{ij}	1.477			

Random Effects	Variance Component	df	χ^2	p-value
u_{0j}	0.360	21	62.08	0.000
u_{1j}	0.724	21	59 .2 7	0.000
r_{ij}	1.469			

Table B.5. Hierarchical Linear Models for Behavioral Management

	Intercepts and Slopes as Outcomes With Season and Grade			Final Model With Additional L-1 Covariates			
Fixed Effect	Coefficient	s. .	p-value	Fixed Effect	Coefficient	s.e.	p-value
Model for Level-1 intercept, β_{0j}				Model for Level-1 intercept, β_{0j}			
Intercept, γ_{00}	4.71	0.18	0.000	Intercept, γ_{00}	4.97	0.20	0.000
Grade 3, γ_{01}	0.08	0.37	0.823	Grade 3, γ_{01}	0.08	0.36	0.82 3
Model for Spring, β_{1j}				Model for Spring, β_{1j}			
Intercept, γ_{10}	0.41	0.23	0 .083	Intercept, γ_{10}	0.39	0.23	0.102
Grade 3, γ_{11}	-0.61	0.40	0.146	Grade 3, γ_{11}	-0 .60	0.40	0.145
				Time			
				Mid-morning, γ_{20}	-0.54	0.12	0.000
				Afternoon, γ_{30}	-0.63	0.12	0.000
				Format			
				Small group, γ_{40}	0.55	0.28	0.048
				Independent Work, γ_{50}	-0.01	0.21	0.964
				Multiple Formats, γ_{60}	0.12	0.19	0.534
				Content area			
				English, γ_{70}			
				Math , γ_{80}			

Random Effects	Variance Component	df	χ^2	p-value
u_{0j}	0.550	21	86.92	0.000
u_{1j}	0.511	21	48.51	0.001
r_{ij}	1.395			

Random Effects	Variance Component	df	χ^2	p-value
u_{0j}	0.511	21	86. 10	0.000
u_{1j}	0.565	21	53.42	0.000
r_{ij}	1.307			

Table B.6. Hierarchical Linear Models Productivity

	Intercepts and Slopes as Outcomes With Season and Grade			Final Model With Additional L-1 Covariates			
Fixed Effects	Coefficient	t	s.e.	p-value	Coefficient	s.e.	p-value
Model for Level-1 intercept, β_{0j}							
Intercept, γ_{00}	4.81	0.16	0.0 00	4.62	0.18	0.000	
Grade 3, γ_{01}	-0.25	0.29	0.395	-0.2 5	0.27	0.362	
Model for Spring, β_{1j}							
Intercept, γ_{10}	0.15	0.24	0.524	0.17	0.23	0.457	
Grade 3, γ_{11}	-0.27	0.37	0.473	-0.26	0.36	0.472	
Time							
Mid-morning, γ_{20}							
Afternoon, γ_{30}							
Format							
Small group, γ_{40}							
Independent Work, γ_{50}							
Multiple Formats, γ_{60}							
Content area							
English, γ_{70}				0.34	0.14	0.017	
Math , γ_{80}				-0.08	0.23	0.736	

Random Effects	Variance Component	df	χ^2	p-value
u_{0j}	0.275	21	51.31	0.000
u_{1j}	0.414	21	42.36	0.004
r_{ij}	1.550			

Random Effects	Variance Component	df	χ^2	p-value
u_{0j}	0.249	21	48.63	0.001
u_{1j}	0.398	21	41.71	0.005
r_{ij}	1.532			

Table B.7. Hierarchical Linear Models for Instructional Learning Formats

	Intercepts and Slopes as Outcomes With Season and Grade			Final Model With Additional L-1 Covariates			
Fixe Effects	Coefficient	s.e.	p-value	Fixed Effects	Coefficient	s.e.	p-value
Model for Level-1 intercept, β_{0j}				Model for Level-1 intercept, β_{0j}			
Intercept, γ_{00}	4. 17	0.16	0.000	Intercept, γ_{00}	3.95	0.20	0.000
Grade 3, γ_{01}	-0.62	0.30	0.048	Grade 3, γ_{01}	-0.40	0.29	0.187
Model for Spring, β_{1j}				Model for Spring, β_{1j}			
Intercept, γ_{10}	-0.37	0.24	0.1 4	Intercept, γ_{10}	-0.20	0.22	0.377
Grade 3, γ_{11}	0.66	0.37	0.091	Grade 3, γ_{11}	0.43	0. 34	0.220
				Time			
				Mid-morning, γ_{20}			
				Afternoon, γ_{30}			
				Format			
				Small group, γ_{40}	-0.21	0.33	0.534
				Independent Work, γ_{50}	-0.99	0.21	0.000
				Multiple Formats, γ_{60}	-0.06	0.21	0.775
				Content area			
				English, γ_{70}	0.36	0.11	0.001
				Math , γ_{80}	0.41	0.19	0.028

Random Effects	Variance Component	df	χ^2	p-value
u_{0j}	0.282	21	52.49	0.000
u_{1j}	0.4 55	21	44.58	0.002
r_{ij}	1.497			

Random Effects	Variance Component	df	χ^2	p-value
u_{0j}	0.304	21	57.04	0 000
u_{1j}	0.379	21	41.38	0.005
r_{ij}	1.397			

Table B.8. Hierarchical Linear Models for Concept Development

	Intercepts and Slopes as Outcomes With Season and Grade			Final Model With Additional L-1 Covariates			
Fixed Effects	Coefficient	s.e.	p-value	Fixed ffects	Coefficient	s.e.	p-value
Model for Level-1 intercept, β_{0j}				Model for Level-1 intercept, β_{0j}			
Intercept, γ_{00}	3.38	0.25	0.000	Intercept, γ_{00}	3.41	0.25	0.000
Grade 3, γ_{01}	-0.77	0.32	0.026	Grade 3, γ_{01}	-0.62	0.32	0.067
Model for Spring, β_{1j}				Model for Spring, β_{1j}			
Intercept, γ_{10}	-0.62	0.25	0.023	Intercept, γ_{10}	-0.54	0.25	0.045
Grade 3, γ_{11}	0.97	0.42	0.032	Grade 3, γ_{11}	0.81	0.37	0.042
				Time			
				Mid-morning, γ_{20}			
				Afternoon, γ_{30}			
				Format			
				Small group, γ_{40}	0.34	0.30	0.251
				Independent Work, γ_{50}	-0.87	0.22	0.000
				Multiple Formats, γ_{60}	0.07	0.14	0.594
				Content area			
				English, γ_{70}			
				Math , γ_{80}			

Random Effects	Variance Component	df	χ	p-value
u_{0j}	0.537	21	79.25	0.000
u_{1j}	0.587	21	49.90	0.001
r_{ij}	1.580			

Random Effects	Vari ance Component	df	χ^2	p-value
u_{0j}	0.549	21	83.63	0.000
u_{1j}	0.450	21	43.59	0.003
r_{ij}	1.501			

Table B.9. Hierarchical Linear Models for Quality of Feedback

	Intercepts and Slopes as Outcomes With Season and Grade			Final Model With Additional L-1 Covariates			
Fixed Effects	Coefficient	s.e.	p-value	Fixe Effects	Coefficient	s.e.	p-value
Model for Level-1 intercept, β_{0j}				Model for Level-1 intercept, β_{0j}			
Intercept, γ_{00}	3.59	0. 28	0.000	Intercept, γ_{00}	3.81	0.30	0.000
Grade 3, γ_{01}	-0.94	0.35	0.014	Grade 3, γ_{01}	-0.88	0.35	0.021
Model for Spring, β_{1j}				Model for Spring, β_{1j}			
Intercept, γ_{10}	-0.49	0.33	0.151	Intercept, γ_{10}	-0.42	0.32	0.211
Grade 3, γ_{11}	0.82	0.41	0.056	Grade 3, γ_{11}	0.71	0.38	0.077
				Time			
				Mid-morning, γ_{20}	0.00	0.17	0.986
				Afternoon, γ_{30}	-0.40	0.19	0.032
				Format			
				Small group, γ_{40}	-0.05	0.42	0.906
				Independent Work, γ_{50}	-0.70	0.25	0.006
				Multiple Formats, γ_{60}	-0.13	0.20	0.510
				Content area			
				English, γ_{70}			
				Math , γ_{80}			

Random Effects	Variance Component	df	χ^2	p value
u_{0j}	0.695	21	95.71	0.000
u_{1j}	0.788	21	59.35	0.000
r_{ij}	1.597			

Random Effects	Variance Component	df	χ^2	p-value
u_{0j}	0.715	21	99.85	0.000
u_{1j}	0.660	21	53.75	0.000
r_{ij}	1.552			

Table B.10. Hierarchical Linear Models for Language Modeling

	Intercepts and Slopes as Outcomes With Season and Grade			Final Model With Additional L-1 Covariates			
Fixed Effects	Coefficient	s.e.	p-value	Fixed Effects	Coeff icient	s.e.	p-value
Model for Level-1 intercept, β_{0j}				Model for Level-1 intercept, β_{0j}			
Intercept, γ_{00}	2.73	0.23	0.000	Intercept, γ_{00}	2.97	0.26	0.000
Grade 3, γ_{01}	-0.22	0.31	0. 490	Grade 3, γ_{01}	-0.10	0.32	0.766
Model for Spring, β_{1j}				Model for Spring, β_{1j}			
Intercept, γ_{10}	-0.41	0.21	0.060	Intercept, γ_{10}	-0.31	0.19	0.121
Grade 3, γ_{11}	0.62	0.32	0.063	Grade 3, γ_{11}	0.46	0.27	0.096
				Time			
				Mid-morning, γ_{20}	-0.23	0.14	0.099
				Afternoon, γ_{30}	-0.36	0.14	0.013
				Format			
				Small group, γ_{40}	0.24	0.36	0.501
				Independent Work, γ_{50}	0.90	0.20	0.000
				Multiple Formats, γ_{60}	-0.11	0.18	0.529
				Content area			
				English, γ_{70}			
				Math , γ_{80}			

Random Effects	Variance Component	df	χ^2	p-value
u_{0j}	0.5 16	21	92.30	0.000
u_{1j}	0.272	21	38.66	0.011
r_{ij}	1.242			

Random Effects	Variance Compo nent	df	χ^2	p-value
u_{0j}	0.532	21	99.52	0.000
u_{1j}	0.147	21	30.54	0.081
r_{ij}	1.556			