Redesigning the Annual NYC School Survey
Improving Measures of School Climate through a Strong Research-Practice Partnership

Part III: How Strong Are the New Measures?

Lisa Merrill
Camille Lafayette

February 2018
### Contents

**Part III: How Strong Are the New Measures?**

<table>
<thead>
<tr>
<th>Topic</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>Defining Survey Measurement Properties</td>
<td>1</td>
</tr>
<tr>
<td>Reliability</td>
<td>1</td>
</tr>
<tr>
<td>Within-School Agreement and Precision</td>
<td>2</td>
</tr>
<tr>
<td>Validity</td>
<td>3</td>
</tr>
<tr>
<td>Assessing the Measures</td>
<td>4</td>
</tr>
<tr>
<td>Effective Leadership</td>
<td>4</td>
</tr>
<tr>
<td>Strong Family-Community Ties</td>
<td>5</td>
</tr>
<tr>
<td>Collaborative Teachers</td>
<td>6</td>
</tr>
<tr>
<td>Supportive Environment</td>
<td>7</td>
</tr>
<tr>
<td>Rigorous Instruction</td>
<td>9</td>
</tr>
<tr>
<td>Trust</td>
<td>10</td>
</tr>
<tr>
<td>Conclusion</td>
<td>11</td>
</tr>
</tbody>
</table>

**References**

12
Redesigning the Annual NYC School Survey: Part III

How Strong Are the New Measures?

As described in previous sections of Redesigning the Annual NYC School Survey, in the summer of 2014, the NYC Department of Education and the Research Alliance for NYC Schools undertook a complete redesign of the district’s annual Survey. The redesign was aimed at capturing better information about schools’ strengths and weaknesses, particularly in terms of the capacities outlined in the City’s Framework for Great Schools.

Below, we describe some criteria to consider when assessing the quality of a survey measure: reliability, within-school agreement, and different types of validity. We then report on the quality of the measures included in the redesigned Survey, organized according to the six elements of the Framework. It is important to note that this information is based on the first administration of the new Survey (in the 2014-2015 school year), and a number of changes were made in response to these findings (changes that appeared on the 2015-2016 Survey and beyond). We plan to continue working with the NYC DOE to analyze and improve the survey measures over time. Below, we highlight measures that met quality standards as of 2014-2015 and provide suggestions to improve the measures that did not.

Defining Survey Measurement Properties

Reliability

In the broadest terms, reliability assesses the extent to which a measure produces consistent results. One way to assess whether a measure is reliable is to see whether individuals tend to respond similarly to all of the items within the measure. This matters because the individual items in a measure are supposed to capture the same underlying concept. Statistically, the more reliable a measure is, the more positively correlated its items are to one another. Researchers commonly use a calculation called Cronbach’s alpha to determine a measure’s reliability (Cronbach, 1951). Alphas
range from 0 to 1, and we use the industry standard of an alpha of .70 or higher as our threshold for sufficient reliability.

**Within-School Agreement and Precision**

The Survey is designed to measure school-level characteristics based on individual respondents’ perspectives. So, it is important to assess how well the survey measures are capturing a common, school-wide characteristic. We do this by calculating how much agreement there is between individuals within the same school. If there are high levels of agreement between different individuals within schools, then they are probably identifying something that is a school-level characteristic. If there are low levels of agreement, then perceptions about that measure vary widely. The degree to which a measure is capturing a school-wide characteristic can be determined using a statistic known as Intra-Class Correlation (ICC). The ICC ranges from 0 to 1, and the higher the number, the more agreement there is within schools. An extreme example of high within-school agreement would be if everyone within each school responded the same way as everyone else in the school; then the ICC would be 1. Likewise, an extreme example of no within-school agreement would be if everyone within each school reported something totally different from one another; then the ICC would be 0. For the purposes of the NYC Survey, we consider within-school agreement to be high if the ICC is above .20, low if it is less than .10, and moderate if it is between .10 and .20 (Raudenbush and Bryk, 2002).

Another consideration of measurement quality related to within-school agreement is school-level precision. The precision of a measure indicates how much error our school-level scores have and is particularly important when using the measures in models to predict other outcomes. The more precise a measure, the better it is at predicting other outcomes we care about. Technically, precision is a function of within-school agreement and the number of surveys per school. The higher the within-school agreement and the larger the number of surveys per school, the greater the precision of a measure. While the ideal measurement properties include high within-school agreement and many respondents per school, even measures with low within-school agreement can be reasonably precise if many people within a school respond to the Survey. For example, many of the student and parent survey measures are precise because there are, on average, hundreds of respondents per school who respond to the Survey. Because there are fewer teachers per school, teacher measures need to have moderate to high within-school agreement to be very precise.
Validity

Validity assesses the extent to which a measure is capturing what it is intended to. There are multiple types of validity, but they generally fall into two large categories: *Construct validity* determines if the items within a measure are asking about the right things in an accessible way, while *criterion validity* determines if the items within a measure are asking about the right topic by calibrating survey measures against a known standard (e.g., other survey measures). Below we provide more information about these types of validity and how we assess them.

**Construct Validity**

- A measure has achieved *face validity* when respondents and other stakeholders read the survey items and agree that they could represent the concept that underlies the measure. We assessed face validity by sharing the Survey with teachers, parents, students, and district employees. When stakeholders agreed that the items represented the concept, we then determined that the measure had face validity. In the future, we could improve our measure of face validity by asking more systematic questions to a formal, pre-specified group of stakeholders.

- A measure has *content validity* when the survey items address the correct components of the measure. We generally assessed the content validity of the survey measures using relevant literature. For measures without much documentation, we used conversations with NYC DOE officials to determine the components of the measures. For example, based on literature about school safety, we determined that the Safety measure should cover physical safety inside and outside the building as well as bullying. We decided that measures with a clear basis in the literature and/or clear definition from the NYC DOE had content validity. Measures without a literature basis or clear definition of components from the NYC DOE had uncertain content validity.

**Criterion Validity**

- A measure has *concurrent validity* when it is positively correlated with another standard at the same point in time, which indicates that the measures are conceptually similar. For example, according to the NYC DOE theory and previous literature, measures within each of the Framework’s elements should be related to current levels of student achievement. To determine concurrent validity, we calculated the correlation between the school-level
average for each measure (across respondent types) with that school’s averages for state ELA and Math test scores (for elementary schools and middle schools) and graduation rates (for high schools).

- A measure has *predictive validity* when it predicts a future outcome that, in theory, it should be able to predict. For example, according to the DOE’s understanding of the Framework for Great Schools, school performance on the Framework elements should predict future levels of student achievement and likelihood of school improvement. We have not yet assessed predictive validity for the new survey measures but plan to, as the next step in our analysis.

### Assessing the Measures

We analyzed these measurement characteristics for each of the elements in the Framework for Great Schools. We provide a discussion of the Survey’s measurement properties below, arranged by element of the Framework. Please see Appendix C for more detailed results of our measurement quality analyses, including statistics about each survey measure for each respondent group.

#### Effective Leadership

The Effective Leadership element consists of five measures: Principal Inclusive Leadership, Teacher Influence, Principal Instructional Leadership, School Leadership Team, and Program Coherence. All five measures were included on the teacher and parent versions of the Survey. All of the measures, with the exception of School Leadership Team (SLT), were previously validated. Overall, the element and the measures performed well on the properties we assessed.

- All measures had high reliability, as indicated by an alpha over .70.
- The within-school agreement, or ICCs, were high for all of the measures, with the exception of SLT.
- All measures—again, with the exception of SLT—had face and content validity. To create the SLT measure, we used documents created by the NYC DOE and asked if the SLT was meeting the requirements outlined by the NYC DOE regulations. However, there may be additional questions about the functioning of the SLT that are important to ask. For example, is the SLT perceived as having a leadership role in the school rather than a bureaucratic committee without much
power? Furthermore, in focus groups, both teachers and parents told us that they did not know enough about the SLTs to answer some of the questions, and many teachers and parents skipped these questions when taking the Survey.

- All measures had concurrent validity (i.e., had positive correlations with student academic achievement). This was particularly true in elementary schools, where these correlations were generally greater than .30.

Based on these findings, we recommended that the NYC DOE reconsider how they capture information about SLTs. We suggested asking if teachers know about what the SLT does rather than asking teachers to judge the quality of the SLT. The NYC DOE accepted our recommendation and removed most of the SLT items for the 2015-2016 survey.

**Strong Family-Community Ties**

The *Strong Family-Community Ties* element consists of two measures: Parent Involvement and Outreach to Parents. The measures are reported by parents and teachers. The teacher measures were derived from the Chicago Consortium’s teacher surveys, and the parent measures were adapted from the teacher measures to be applicable to parents.

- All measures had high reliability, as indicated with an alpha over .70.
- All measures had a moderate amount of within-school agreement (i.e., an ICC between .1 and .2).
- The content validity was uncertain. When developing the measures, we found that they were often difficult to define because the term “community” is not clearly defined. In NYC, students do not necessarily attend schools in their own neighborhood. Therefore, it is difficult to know if the “community” refers to the geographic area around the school, the neighborhood the school is in, the neighborhoods that the students and their families live in, or the people connected to the school (the students, their families, and school staff). In terms of family involvement, it is also unclear which kind of involvement is most important to the NYC DOE, and we had few previously validated measures to work with. Therefore, we do not know if we identified the correct dimensions of Family and Community Ties. Parent focus groups provided some evidence that the face validity for these measures was questionable. Parents were confused by the selection of activities that were associated with parental involvement. For
example, were parents supposed to be volunteering in their schools? If so, how often?

- The concurrent validity was mixed. It was strong for teachers, particularly in elementary school, and weak for parents. The teacher measures were positively associated with student outcomes and the parent measures were negatively associated with student outcomes.

We initially recommended modifying the scales of these measures. However, upon further reflection and analysis, we recommended developing a clearer vision for parent and community ties that is appropriate for New York City’s schools. The NYC DOE accepted our recommendation and revised items within the Outreach to Parents and Parent Involvement measures on both the parent and teacher versions of the Survey. Preliminary analysis of the 2015-2016 survey shows that these measures are much improved but may still need some tweaking.

**Collaborative Teachers**

The Collaborative Teachers element is made up of nine measures: Peer Collaboration, Collective Responsibility, Reflective Dialogue, Cultural Awareness, Inclusive Classroom Instruction, Innovation, School Commitment, and Quality Professional Development. The element draws from each respondent group (i.e., teachers, parents and students). All of these are from previously validated measures. Overall, the measures were of good, but varied, quality.

- All measures have high reliability, with an alpha over .70.
- Many measures had at least a moderate level of within-school agreement. However, the Cultural Awareness, Inclusive Classroom Practice, and Reflective Dialogue measures had low levels of within-school agreement. School Commitment and Focus on Student Learning had high within-school agreement, and the rest had a moderate amount.
- All measures have face and content validity. District personnel and teachers thought that these questions covered the appropriate topics for the measures.
- All measures except Cultural Awareness (student and parent) and Quality Professional Development (high school) had concurrent validity, as indicated by positive correlations with academic achievement. The correlations were strongest for elementary school teachers.
Our analysis also suggested that some measures within this element might be redundant with one another. In several cases, we found that two measures had near perfect concurrent validity with one another, meaning they were measuring nearly the same concept. As a result, we suggested removing the Inclusive Classroom, Peer Collaboration, and Focus on Student Learning measures from the next year’s survey to reduce its length. In addition, in correlating Cultural Awareness (student and parent) with academic achievement, we found the relationship to be negative in elementary and middle school, indicating poor concurrent validity. These findings were also supported by our focus groups with students—they reported that the items were confusing and challenging to answer. Quality Professional Development also had a weak relationship with academic achievement for high schools. However, at the elementary and middle school level, we found the relationship to be positive and significant.

The NYC DOE accepted our recommendation and removed the Focus on Student Learning measures from the teacher survey and elected to modify the Peer Collaboration and Inclusive Classroom measures. The NYC DOE decided to collapse Inclusive Classroom into Cultural Awareness, refine the language used to formulate one measure, and rephrase one of the items within Peer Collaboration. Preliminary analysis shows that Cultural Awareness and Peer Collaboration are much improved—with fewer missing answers, higher averages, and strong correlations with academic achievement—but may still need some wording adjustments (based on more missing responses than we would like). Overall, as a result of the revisions, the element has been streamlined while maintaining its high reliability and within-school agreement.

**Supportive Environment**

The Supportive Environment element consists of eight measures: Safety, Social Emotional, Academic Press, Peer Interactions, Classroom Behavior, Academic Personalism, Peer Support for Academic Work, and Next-Level Guidance. These measures were previously validated, except for Social Emotional Measure and Next-Level Guidance. For middle and high schools, all of the measures are reported by students. Elementary school students do not take the Survey, so teachers report on the measures.

- The student measures had high reliability, with alphas above .70.
- The within-school agreement of these measures was low for all measures with the exception of Safety, particularly when the measures were based on student
responses. This could suggest that students within a school experience some aspects of the Supportive Environment measure differently, but have more similar feelings about safety.

- The face and content validity were strong for most measures, except for Social Emotional Measure and Next-Level Guidance.
  - Teaching social and emotional skills is complicated, and the working group did not have a clear understanding of the particular instructional strategies or techniques for teaching social and emotional skills that the NYC DOE wanted to capture. Ultimately, we designed a measure that addressed whether teachers focused on the social and emotional skills that the NYC DOE believes are particularly important—perseverance, self-confidence, critical thinking, and self-advocacy—rather than how they taught these skills. With more time, we would have drawn more on additional research about successful practices for teaching social and emotional skills.
  - We also found the Next-Level Guidance measure to be conceptually vague. It was unclear if the measure should focus on helping prepare students for their next stage in education or their next stage in life, so we included items about both in the measure. For high school students, we asked about college, career, and life skills. For middle school students, we asked about various aspects of the high school application process and life skills.

- All measures had concurrent validity and positive correlations with student academic achievement, particularly for elementary school, except for Academic Personalism (student), which did not have a significant relationship with academic achievement.

We recommended that the NYC DOE reconfigure its Social Emotional and Next-Level Guidance measures (and possibly consider removing the Social Emotional Measure altogether). The NYC DOE accepted our recommendation and revised both the Next-Level Guidance and Social Emotional Measure on the teacher survey to be more pointed and specific. The NYC DOE also revised the Next-Level Guidance measure on the high school student version of the Survey to separate items into two groups: one that asks about college, and one that asks about career preparation. Preliminary analysis of the 2015-2016 survey shows that these measures are much improved on both the teacher and student versions of the Survey.
Rigorous Instruction

The Rigorous Instruction element consists of three measures: Common Core Shifts in Literacy and Math, Quality of Student Discussion, and Course Clarity. Common Core Shifts in Literacy and Math were based on U.S. Department of Education implementation surveys about instructional shifts of the Common Core. Quality of Student Discussion and Course Clarity were developed and validated by the Chicago Consortium. All of these measures are on the teacher version of the Survey, except for Course Clarity, which is on the student survey.

- The two measures from Chicago (Quality of Student Discussion and Course Clarity) have high reliability; all measures have alphas above .70.
- Two of the four measures had very low within-school agreement: Common Core Shifts in Literacy and Common Core Shifts in Math. Research has long documented that there is great variety of teaching styles and quality within schools (Rowan et al., 2004; Weick, 1976). It is possible that instruction is less of a school-level phenomenon than other aspects of the Framework. At the same time, it could also be that measuring instruction via a teacher self-report survey is not the most accurate means of assessment. The student measure of instruction, Course Clarity, also had low within-school agreement. The measure with the highest level of within-school agreement was Quality of Student Discussion.
- All measures had face and content validity. The measures were based on prior research on effective instruction and the major instructional shifts required by the Common Core.
- All measures had concurrent validity and had positive correlations with student academic achievement. The correlation was particularly strong for elementary school teachers.

We recommended that the NYC DOE reconsider the Common Core Shifts in Literacy and Common Core Shifts in Math measures. The NYC DOE accepted our recommendation and changed the aim of both measures. They elected to ask about resources and tools available to support teachers and changed the scale from a frequency scale to an agreement scale. Preliminary analysis shows that these measures are much improved on the 2015-16 teacher survey.
Trust

The Trust element consists of five measures: Parent-Teacher Trust, Parent-Principal Trust, Student-Teacher Trust, Teacher-Teacher Trust, and Teacher-Principal Trust. The element includes responses from parents, students, and teachers.

- All measures have high reliability; all had alphas above .70.
- The within-school agreement varied. The measures reported by teachers (Teacher-Teacher Trust and Teacher-Principal Trust) had higher degrees of within-school agreement than the student or parent measures. In particular, the Teacher-Principal measure has a high amount of within-school agreement. The Parent-Principal Trust measure had the lowest level of within-school agreement among all the measures in this element.
- All measures had face and content validity. The district personnel and teachers thought that these questions covered the appropriate topics for the measures.
- All measures had concurrent validity and positive correlations with student academic achievement.

We proposed removing a few items from the measures in this element, simply to make the Survey shorter. The NYC DOE accepted our recommendation and elected to remove a few items and provide more specificity to select items on all five Trust measures. Preliminary analysis of the 2015-2016 survey shows that these changes improved the measures within this element. Based on feedback from the student focus groups, we also suggested adding a Student-Principal Trust Measure. However, the NYC DOE did not add this new measure.
Conclusion

Overall, the measures used on the NYC School Survey to assess the six elements of the Framework for Great Schools perform well. They were all reliable and, for the most part, demonstrated face, criterion, and concurrent validity. Only four of 32 measures did not have concurrent validity, and three of these did not demonstrate content validity either. Two additional measures had neither content nor face validity. We recommended changes to all six of these measures, and early analyses show that the measures were improved on the 2015-16 surveys.

The amount of within-school agreement varied by element and respondent group. As described above, we made some recommendations for revising some measures, in an effort to improve within-school agreement. At the same time, some measures may be attempting to capture aspects of schools that individual students, teachers, and parents perceive differently. For example, we found low within-school agreement on the Student-Teacher Trust measure. It is possible that the measures should be re-written, but it is also possible that the measure is accurately capturing the fact that, within a single school, different students have very different levels of trust in their teachers. This latter theory suggests that we do not have to re-write those measures but instead should pay attention to both the variation between schools as well as within schools. We will continue to explore these possibilities in future work.

Check out the rest of our Redesigning the Annual School Survey collection.
References


