Journal of Applied Developmental Psychology 55 (2018) 119–129

Contents lists available at ScienceDirect

Journal of Applied Developmental Psychology

journal homepage: www.elsevier.com/locate/jappdp

Development and implementation of student social-emotional surveys in the CORE Districts

Martin R. West[a,*], Katie Buckley[b], Sara Bartolino Krachman[b], Noah Bookman[c]

[a] CORE Districts, 1107 9th Street, Suite 500, Sacramento, CA 95814, United States
[b] Transforming Education, 24 School Street, 3rd Floor, Boston, MA 02108, United States
[c] Harvard University, Harvard Graduate School of Education, 6 Appian Way, Gutman Library 454, Cambridge, MA 02138, United States

ARTICLE INFO

Keywords:
CORE Districts
Social-emotional learning
Student surveys
Survey validation

ABSTRACT

States and school districts across the U.S. are seeking to expand their definition of student success to include social-emotional learning. The CORE Districts, a collaborative of California districts that has developed a system of school accountability and continuous improvement that includes measures of social-emotional skills based on student self-reports, exemplify this trend. In this case study, we provide an overview of CORE's School Quality Improvement System, which was implemented in the 2015–16 school year across six districts serving roughly one million students. We describe how four social-emotional competencies—growth mindset, self-efficacy, self-management, and social awareness—were selected for assessment; the process for curating and piloting assessments of students' social-emotional skills; and reliability and validity evidence from a 2015 field test of social-emotional measures based on self-reports from nearly 400,000 students. We conclude with lessons from the development of CORE's system for other next-generation accountability and continuous improvement efforts.

The CORE Districts (or CORE) is a partnership of California local educational agencies working to improve student achievement by fostering collaboration and learning across its eight members: Fresno, Garden Grove, Long Beach, Los Angeles, Oakland, Sacramento, San Francisco, and Santa Ana Unified School Districts. CORE's governing board comprises the superintendents of its member districts, and administrators, school leaders, and teachers from each district are actively involved in collaborative activities. In 2013, CORE applied for and received a waiver from the U.S. Department of Education that provided six of its member districts flexibility from key requirements of the school accountability system prescribed by the federal No Child Left Behind Act. Through this waiver, CORE sought to implement a new type of accountability system that, rather than looking solely at test scores and graduation rates, incorporated schools' performance across a broader range of outcome measures.

In particular, the CORE governing board wanted to include measures of social-emotional (SE) skills and school culture/climate (CC), alongside traditional academic indicators, in a more holistic index of school quality. They focused on SE skills because of (a) research demonstrating their importance for students' academic, career, and life success (e.g., Almlund et al., 2011; Moffitt et al., 2011; Heckman, Stixrud, Urzua, 2006); (b) benefits two member districts had seen from implementing social-emotional learning (SEL) programs as part of the Collaborative for Academic, Social, and Emotional Learning's Collaborating Districts Initiative (Kendziora & Osher, 2016); and (c) a shared recognition that the development of SE skills was largely missing from districts' existing performance measurement systems.

In this paper, we provide an overview of CORE's School Quality Improvement System, which was fully implemented in the 2015–16 academic year. We describe how four SE competencies—growth mindset, self-efficacy, self-management, and social awareness—were selected for assessment. We discuss the process for curating and piloting survey-based measures of these competencies, and present validity and reliability evidence from a 2015 field test involving nearly 400,000 students. We conclude with lessons from the development and early implementation of CORE's system that can inform other next-generation assessment and continuous improvement efforts.

1. CORE's School Quality Improvement System

The U.S. Department of Education approved CORE's NCLB waiver application in August 2013, authorizing the development and implementation of its proposed School Quality Improvement System (SQIS). The key principles of the SQIS are captured in the four

* Corresponding author.
E-mail addresses: martin_west@gse.harvard.edu (M.R. West), katie@transformingeducation.org (K. Buckley), sara@transformingeducation.org (S.B. Krachman), noah@coredistricts.org (N. Bookman).

http://dx.doi.org/10.1016/j.appdev.2017.06.001
Received 25 September 2016; Received in revised form 11 May 2017; Accepted 1 June 2017
Available online 14 July 2017
0193-3973/

"foundational goals" articulated in its waiver request:

1. College- and career-ready expectations for *all* students.
2. A focus on collective responsibility, accountability, and action that emphasizes capacity-building over accountability.
3. The development of intrinsic motivation for change through differentiated recognition, accountability, and support for schools.
4. Focused capacity-building for effective instruction and leadership.

The first goal emphasizes high standards for all students, coupled with a commitment to eliminate outcome disparities across student subgroups. Rather than mandating how to accomplish this, however, CORE tasks individual districts with making their own instructional and programming decisions in order to preserve local autonomy. The second and third goals reflect a focus on cultivating a sense of collective responsibility among educators for students' success, rather than relying on the type of punitive sanctions that characterize some educational accountability systems. Pursuant to the fourth goal, capacity-building occurs by forming Communities of Practice and by matching higher- and lower-performing schools in School Pairings to support continuous improvement. CORE also offers professional development, tools, research, and convenings to enable district leaders and educators to share best practices.

CORE's School Quality Improvement Index (SQII or the "Index") serves as the foundation of the SQIS. While SQIS refers to the full system of accountability and continuous improvement, the Index is the quantitative formula used to assess school performance. The Index includes measures of student academic achievement and growth, student social-emotional competencies, and school culture/climate in order to provide a holistic picture of student success and school quality.

The Index has been rolled out in stages over the course of three years. Fig. 1 shows the full Index as implemented for the first time in the 2015–16 academic year. Academic indicators account for 60% of the Index, while social-emotional and school culture/climate factors account for 40%. Consistent with its commitment to continuous improvement, CORE plans to revise the indicators included in the Index and the weights assigned to them over time based on feedback from stakeholders and developments in research.

2. Selecting social-emotional competencies to assess

While CORE committed in its waiver to measure students' social-emotional development, it did not identify the specific competencies or assessments the districts would use. Accordingly, one of their first steps after the waiver's approval was to select an initial set of social-emotional competencies for assessment and identify promising measures for each. In November 2013, the governing board convened SEL experts and representatives from each CORE District including superintendents, directors of student support, directors of SEL, and directors of special education. The SEL experts in attendance were from the Collaborative for Academic, Social and Emotional Learning (CASEL), the John W. Gardner Center for Youth at Stanford, and Transforming Education (TransformEd). Adopting a framework proposed by TransformEd known as the "3 Ms," CORE agreed to select specific social-emotional competencies for inclusion in the Index based on the extent to which they met three criteria:

1. *Meaningful* indicates that the competency predicts important academic, career, or life outcomes.
2. *Measureable* indicates that the competency can be measured reliably through a valid assessment that is feasible to administer at scale in schools.
3. *Malleable* indicates that the development of the competency can be influenced in an educational setting.

In order to ensure that the assessment of multiple SE competencies would yield data that were complementary rather than redundant, CORE additionally decided to include at least one *intra*personal skill and one *inter*personal skill among its initial set of SE competencies (National Research Council, 2012). District representatives and SEL content experts used a voting process to identify four specific SE competencies for inclusion in the Index: growth mindset (Dweck, 2006), self-efficacy (Bandura, 1997), self-management (CASEL, 2005), and social awareness (CASEL, 2005). (See Fig. 2 for definitions.) Participants acknowledged that this is not a comprehensive set of SE competencies, but rather a starting point that could be revised as new research emerges. For example, they considered incorporating collaborative problem-solving into the initial set of competencies but CORE elected to wait until performance-based measures of this competency had been piloted in the 2015 Program for International Student Assessment study.

3. Moving from competencies to validated measures

Once the four SE competencies were identified, TransformEd

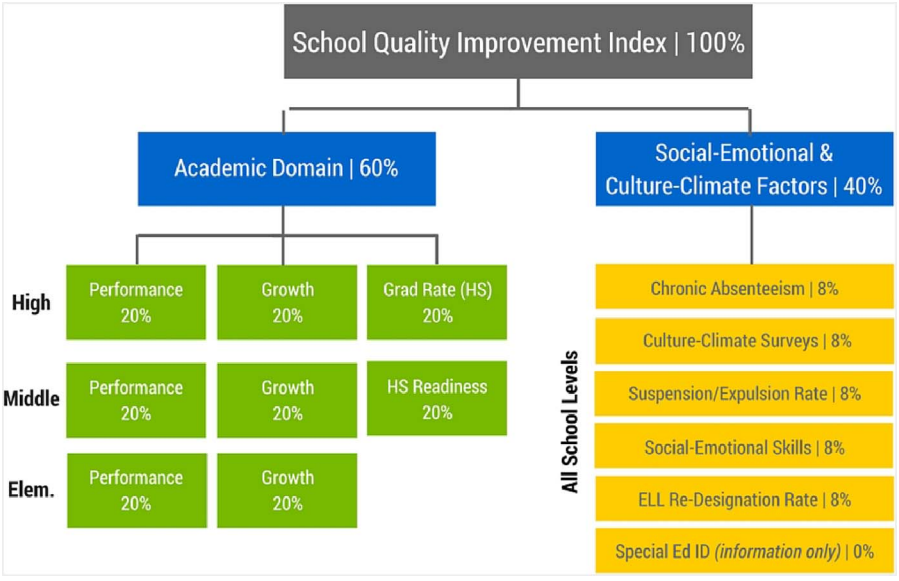

Fig. 1. The School Quality Improvement Index.

Growth mindset:	The belief that one's abilities can grow with effort. Students with a growth mindset see effort as necessary for success, embrace challenges, learn from criticism, and persist in the face of setbacks (Dweck, 2006).
Self-efficacy:	The belief in one's own ability to succeed in achieving an outcome or reaching a goal. Self-efficacy reflects confidence in the ability to exert control over one's motivation, behavior, and environment (Bandura, 1997).
Self-management:	The ability to regulate one's emotions, thoughts, and behaviors effectively in different situations. This includes managing stress, delaying gratification, motivating oneself, and setting and working toward personal and academic goals (CASEL, 2005).
Social awareness:	The ability to take the perspective of and empathize with others from diverse backgrounds and cultures, to understand social and ethical norms for behavior, and to recognize family, school, and community resources and supports (CASEL, 2005).

Fig. 2. Definitions of the social-emotional competencies assessed by CORE.

scanned the field to identify a set of measures that met criteria established by CORE and described below. Despite their limitations (Duckworth and Yeager, 2015; West, 2016), self- and teacher-report surveys were identified as the best available measurement approach. CORE then partnered with the Center for Education Policy Research (CEPR) at Harvard University to analyze the measures' performance in an initial 2014 pilot involving 18 schools and in a broader 2015 field test to validate the measures prior to the full-scale rollout. The results of these analyses informed the final choice of measures and the decision to move forward with the plan to include SE measures in the SQII in the 2015–16 school year.

3.1. Criteria for measures

Core established a set of five criteria to guide the selection of measures of student SE competencies.

3.1.1. Evidence-based

The measures needed to demonstrate emerging evidence of validity and reliability when administered to students in grades 4–12. They also needed to have "face validity" for educators, such that educators believe the scales capture the relevant competency.

3.1.2. Free-to-administer

To ensure CORE's financial sustainability, the measures needed to be free. While the obtaining permission from researchers to use measures they had developed was time-consuming, this was preferable to using third-party assessment providers that would charge a student- or school-based licensing fee each year.

3.1.3. Practical

The measures needed to be easy to administer to students in grades 4–12, with as little administrative burden as possible. Districts asked that there be both online and paper-based survey options given limitations of some schools' technological infrastructure. Further, the districts asked that they be able to incorporate the measures into existing surveys to avoid administering a separate assessment.

3.1.4. Parsimonious

The assessments needed to use the fewest number of items possible to get valid, reliable results. This criteria was adopted in response to concerns about the "over-testing" of students and to teachers' and administrators' desire to protect instructional time. CORE aimed for students to be able to complete SE and CC survey items in approximately 10 min each, with an understanding that proxy indicators such as attendance, grades, and discipline rates used to round out the picture of students' SE skills.

3.1.5. Strengths-based

CORE administrators expressed a preference that questions be phrased in a positive manner whenever possible, unless research showed that a particular negatively-phrased scale was more valid or reliable. For example, they preferred items such as "I can improve my intelligence by working hard" to those such as "My intelligence is something that I can't change very much."

3.2. Measure selection

Based on these criteria, TransformEd sought guidance from experts in SE assessment and curated a set of student self-report surveys to assess each competency. Student reports were developed for grades 4–12 only, given that surveys administered to younger students often require different wording or other accommodations. In addition, teacher reports were developed for the two SE competencies considered to be potentially externally observable (i.e., self-management and social-awareness), both as a potential complement to student reports for inclusion in the Index and as a way to provide information on the social-emotional development of younger students. The extent to which social awareness is in fact externally observable is unclear, as the construct includes students' abilities to empathize with others. As operationalized by CORE, however, the construct is a hybrid of social awareness and social competence, and the questionnaire developed for teachers focused on the latter.

Once draft scales were developed, district staff members vetted the measures and provided feedback focused on their face validity among educators. The CORE Board approved the final set of measures in December 2013, requesting that they be piloted with a small group of schools in spring 2014 and field tested with all schools in spring 2015 before being included in the SQII the following year. Teacher reports of students SE skills were included in the pilot data collection and field-tested in two districts. As discussed below, however, they are not being used in all districts and therefore do not factor into schools' Index scores.

3.3. Minimizing bias

Prior to the pilot data collection, CORE took steps to address three common sources of measurement error in self-report survey responses: social desirability bias, stereotype threat, and reference bias.

3.3.1. Social desirability bias

Social desirability bias refers to the tendency for survey respondents to provide answers that they believe are socially acceptable rather than those that reflect their true thoughts or feelings (Podsakoff, MacKenzie, & Lee, 2003; Fisher, 1993). For example, if a survey asks a student how often she is polite to adults, the student may answer "almost all the

time" even if the response "almost never" more accurately reflects her behavior, simply because she knows that being polite to adults is socially desirable. CORE sought to mitigate this bias in two ways: (a) by explicitly stating that students' survey responses would remain confidential and not influence their grades or be used as a measure of performance; and (b) by asking the adults proctoring survey administration to stand at the back of the classroom instead of circulating.

3.3.2. Stereotype threat

Stereotype threat refers to the tendency for individuals' survey responses to be influenced by their perception of how members of a group with which they identify (e.g., race, ethnicity, gender, or socioeconomic status) are believed by others to perform in the relevant domain (Spencer, Steele and Quinn, 1999). Research on stereotype threat suggests that students who are asked to report their gender or ethnicity before completing an assessment are more likely to perform in a manner consistent with their perception of how people in their identify group are expected to perform (Walton and Spencer, 2009). CORE therefore included demographic questions only at the end of the survey and, in some districts, removed them and relied instead on a bar code as a student identifier that could be linked to data from the districts' student information system.

3.3.3. Reference bias

Reference bias refers to the tendency for individuals' survey responses to be influenced by differing implicit standards of comparison (Heine, Lehman, Peng, & Greenholtz, 2002). For example, when asked to evaluate their self-management skills, students with high expectations for their behavior or work ethic may assign themselves lower ratings than students with lower expectations, even if their actual capacity to regulate their behavior is the same. Of particular concern for the use self-report surveys for the purpose of school accountability is the possibility that a school's culture might influence its students' frames of reference, in turn causing them to interpret items differently than students from other schools (Goldman, 2006; West, Kraft, et al., 2016). To the extent that students attending schools with more demanding expectations for student behavior hold themselves to a higher standard when completing questionnaires, reference bias could make comparisons of responses across schools misleading.

To address concerns about reference bias, CORE partnered with Educational Testing Service to develop anchoring vignettes for self-management and social awareness to be incorporated into its self-report surveys (King et al., 2004). Anchoring vignettes are brief descriptions of hypothetical individuals who exhibit varying levels of a given construct. Respondents are asked to read the vignettes and assign each one a rating using the same response options used to assess themselves. These ratings then provide a basis for re-scaling respondents' self-reports relative to common points of reference. Kyllonen and Bertling (2013) show that adjustments based on anchoring vignettes can reduce the influence of reference bias in international comparisons of student attitudes, strengthening within-country correlations between measures of achievement and academic self-efficacy and reversing paradoxical cross-country correlations that suggest, for example, that students who are less confident in their abilities in math and science achieve at higher levels. To our knowledge, however, anchoring vignettes have not previously been used when assessing students' SE skills in the U.S.

3.4. Pilot testing

CORE conducted a pilot test of the SE measures in spring 2014 with approximately 9000 students and 300 teachers in 18 schools. During the pilot, two different forms of the student self-report surveys and teacher surveys were randomly assigned to participants. For each competency, one of the forms used the original survey scale developed by a contributing researcher and the other used a modified version developed in partnership with educational psychologist Hunter Gehlbach to reflect best practices in survey design (e.g., removing double-barreled items, translating agree/disagree statements into questions). The original scales varied in length; those for growth mindset and self-efficacy had four items each, while those for social awareness and self-management had eight and nine, respectively.

Upon completion of the pilot test, CEPR researchers compared the two scales developed for each construct to identify the more promising one based on (a) predictive validity, including correlations with academic and behavioral indicators and with scales measuring related SE constructs that were included for this purpose; and (b) internal reliability, or the degree to which the individual items included in each scale assessed the same underlying construct. The scale that demonstrated the strongest student-level correlations with theoretically related academic and behavioral outcomes (e.g., grades, attendance, suspensions) and met commonly accepted reliability standards (i.e., Cronbach's alpha > 0.70) was selected for use in the 2015 field test.

Analyses reported elsewhere (West, Dow, & Buckley 2017) indicated that using anchoring vignettes to rescale students' self-reports improved neither the internal reliability of the survey scales nor their correlations with academic and behavioral indicators. These correlations fell in most cases, perhaps due to the introduction of additional measurement error. This could indicate that reference bias is not an important phenomenon for the comparability of student responses within the CORE Districts, or that the particular anchoring vignettes used in the pilot were ineffective in addressing it. Based on these results and the implications of including anchoring vignettes for survey length, CORE elected not to use them during the subsequent field test. Understanding the extent and implications of reference bias nonetheless remains a priority for CORE and its research partners.

4. Field test results: validity and reliability

CORE conducted its field test of the refined student survey measures in spring 2015 across all of the roughly 1500 schools in its member districts, providing an opportunity to analyze the measures' reliability and validity when administered at scale. CORE had committed to incorporating SE data for students in grades 5–12 into the SQII and planned to base its decision on whether or not to include students in grade 4 on the results of the field test. All students in grades 4–12 who were present when the surveys were administered therefore participated in the data collection; one district also administered the surveys to students in grade 3.

4.1. Analytic samples

We focus our analysis of the field test data on 378,465 students across five districts for which individual students' survey responses could be linked to administrative data from the 2014–15 school year acquired by the John W. Gardner Center at Stanford University. These administrative data include demographic information for the full sample and, for various subsamples, grade point averages (GPA), state test scores in English language arts (ELA) and math, attendance, and suspensions. We standardize the test scores within grade and subject to have a mean of zero and a standard deviation of one across all students in the sample for whom scores are available. GPA, absences, and suspensions are standardized by district due to variation in reporting practices.

Table 1 presents the demographic characteristics of the students in our analytic sample. The first column describes the full sample; 68% of these students are eligible for a free or reduced-price lunch, roughly 70% are Hispanic, and 20% are English language learners. The remaining columns describe the subsamples for which we can assess predictive validity based on various outcomes, which range in size from 86,012 students (absences) to 251,672 (GPA). Due to the scale of the data collection, virtually all of the differences in mean characteristics between the full sample and these subsamples are statistically

Table 1
Mean demographics of the CORE field test sample.

Variable	Full sample	GPA sample	Test score sample	Suspensions sample	Absences sample
Male	0.50	0.50	0.50	0.50	0.50
Free/reduced-price lunch	0.68	0.67	0.69	0.68	0.71
English language learner	0.21	0.18	0.22	0.19	0.29
Special education	0.10	0.10	0.10	0.10	0.08
African-American	0.08	0.08	0.08	0.07	0.09
Asian	0.10	0.11	0.10	0.07	0.10
Hispanic	0.71	0.70	0.70	0.76	0.67
White	0.10	0.09	0.11	0.09	0.10
Number of students	378,456	251,672	246,489	213,554	86,012
Number of schools	1106	470	986	430	305

Note. Data from schools with fewer than 25 respondents are removed from the sample. Full sample includes all students whose survey responses could be linked to administrative demographic data; additional samples are restricted to students with non-missing information on GPA, Math and English language arts test scores, suspensions, and absences. All subsample means except underlined values differ significantly from the mean for the full sample ($p < 0.01$).

significant. The students for whom we observe suspensions are five percentage points more likely to be Hispanic, while those with absences are eight percentage points more likely to be English language learners. The latter difference reflects the fact that absences were only available for students in three of the five districts. All other differences are quite small in magnitude, however, at less than three percentage points.

4.2. Reliability

Consistent with CORE's decision-making process, we use Cronbach's alpha as our primary metric to assess the internal reliability of the survey scales used to measure each SE construct. In addition, we present data on the temporal stability of the SE measures when administered to the same student at an interval of one year. Because CORE plans to use the SE data as an indicator of school performance, we also discuss evidence on the extent to which student responses are correlated by school as a measure of school-level reliability.

Table 2 reports the mean, standard deviation, and Cronbach's alpha for each of the four constructs, as well as for an equally weighted average of students' scores on the four separate scales. The first column displays these metrics for the full sample, while the remaining columns focus on specific student subgroups.

The measures generally demonstrate strong internal reliability. For the full sample, the scales used for three of the constructs have alphas considerably above the 0.70 benchmark commonly used as a threshold for evaluating the internal consistency of survey scales. The exception is the scale used to assess growth mindset, for which the alpha was 0.70 exactly, suggesting the value of continued efforts to enhance measurement of this construct.

There are also only minor differences in the measures' reliability across subgroups. While tests based on Feldt (1969) indicate that all of the differences in reliability across groups reported in Table 2 are statistically significant, most are trivial in magnitude. Of particular interest is the finding that the reliabilities are only modestly lower for English language learners and students with disabilities (with differences in the range of 0.01 to 0.05), suggesting that students in these groups did not struggle to interpret the vocabulary used in the survey items substantially more than other students. At the same time, reliabilities for students in grade 3 are lower on each construct than for students in grade 4, for whom the measures consistently exhibit lower reliabilities than for any other subgroup. This relationship between survey reliability and grade level is particularly pronounced for growth mindset, where the estimated reliability for grade 3 students was only 0.60. Based in part on these results, CORE elected to move forward with its plans to incorporate grade 4 into the SQII, effectively doubling the number of students whose responses would be included in the Index for elementary schools ending in grade 5, but not to extend SE data collection to grade 3.

In addition to internal consistency as gauged by Cronbach's alpha, another important aspect of reliability is temporal stability. In an effort to minimize disruption to instructional time, CORE elected not to administer its SE measures to the same students multiple times during the field test in a manner that would make it possible to calculate two-week test-retest reliabilities. We can, however, combine data from the 2015 field test with data from the 2016 implementation of the SQII to provide initial evidence on the measures' temporal stability over one year.

Fig. 3 reports Pearson correlation coefficients for each SE measure and, for purposes of comparison, state test scores through grade 8. As expected, students' math and English language arts test scores are

Table 2
Means, standard deviations, and reliability coefficients for student social-emotional measures, overall and by student subgroup.

		Full sample	Male	ELL	FRPL	SPED	Black	Asian	Hisp	White	Grade 3	Grade 4
Self-management (9 items)	Mean	4.05	3.95	3.86	4.01	3.77	3.93	4.22	4.01	4.28	3.92	4.01
	SD	0.69	0.71	0.74	0.7	0.77	0.72	0.59	0.7	0.61	0.73	0.72
	Alpha	0.85	0.85	0.83	0.85	0.84	0.84	0.84	0.85	0.84	0.80	0.83
Growth mindset (4 items)	Mean	3.73	3.72	3.39	3.67	3.41	3.8	3.85	3.67	4.01	3.58	3.62
	SD	0.96	0.97	0.97	0.96	0.98	0.99	0.9	0.97	0.91	1.04	0.99
	Alpha	0.70	0.70	0.65	0.69	0.66	0.67	0.72	0.69	0.70	0.60	0.62
Self-efficacy (4 items)	Mean	3.48	3.53	3.34	3.45	3.25	3.57	3.56	3.41	3.77	3.69	3.67
	SD	1.00	0.99	1.00	1.00	1.02	1.03	0.94	1.01	0.98	1.00	1.00
	Alpha	0.86	0.86	0.82	0.86	0.81	0.85	0.88	0.86	0.88	0.78	0.82
Social awareness (8 items)	Mean	3.71	3.63	3.67	3.7	3.55	3.64	3.76	3.69	3.88	3.97	3.90
	SD	0.71	0.73	0.74	0.71	0.79	0.77	0.62	0.72	0.66	0.72	0.7
	Alpha	0.81	0.81	0.80	0.80	0.80	0.80	0.80	0.81	0.81	0.76	0.78
	N	378,456	181,060	76,168	246,339	36,507	30,358	37,093	267,645	36,981	32,991	40,747

Note. Table reports the mean, standard deviation (SD), and Cronbach's alpha (Alpha) for each survey scale for the full sample and various subgroups. ELL: English language learner; SPED: special education; FRPL: free/reduced-price lunch; Hisp: Hispanic. All differences in means, SDs, and Alpha coefficients across subgroups are statistically significant ($p < 0.01$).

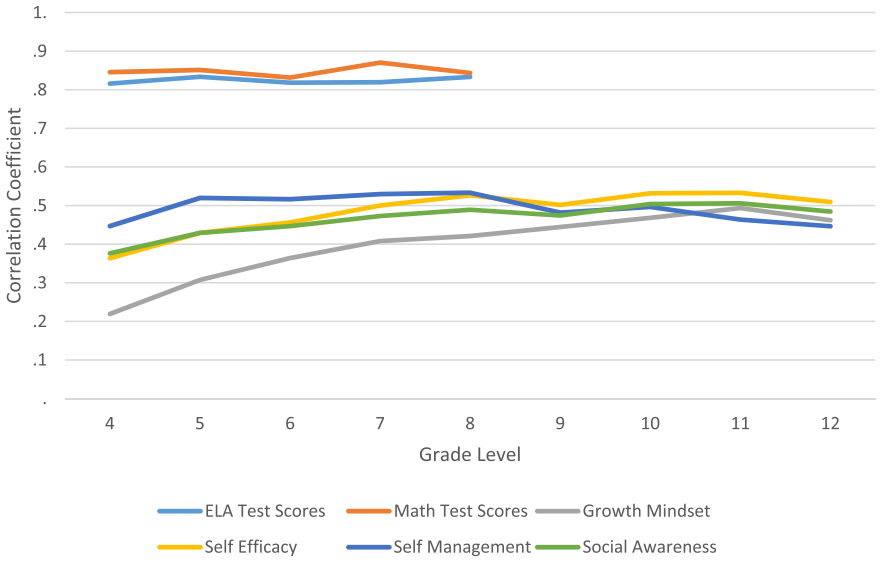

Fig. 3. One-year test-retest reliabilities for academic and social-emotional skills. This figure displays Pearson correlation coefficients for students' English language arts (ELA) and Math test scores and social-emotional skills as assessed in spring 2015 and 2016. Grade level refers to students' grade in spring 2016. Sample size ranges from 22,082 to 42,565 students depending on grade and construct. All reported coefficients are statistically significant at $p < 0.001$.

highly correlated from one year to the next ($r = 0.82–0.87$, $p < 0.001$), and the strength of this relationship is similar across grade levels. Although also statistically significant ($p < 0.001$), the parallel reliability estimates for the SE measures are markedly lower ($r = 0.22–0.53$) and, with the exception of self-management, tend to increase across grades. Additional analyses confirm that adjusting the correlations reported in Fig. 3 for the SE measures' lower internal reliability accounts for only a small fraction of the differences in results. While striking, the lower temporal stability observed for the SE measures is not necessarily a concern given that one reason for educators' interest in SE skills is the notion that they may be more malleable over time than cognitive ability (see, e.g., Almlund et al., 2011).

Because CORE intends to use the SE measures primarily as an indicator of school performance, it is also important to assess the extent to which the measures differ systematically across schools. Hough et al. (2017) draw on the same field test data used in this paper to calculate the intra-class correlation coefficient for the SE measures as a measure of the share of the overall variation in students' scores that is explained at the school level. They find that schools explain only 8%, 7%, and 3% of the variation, respectively, at the elementary, middle, and high school levels. For math test scores, in contrast, schools explain 15–20% of the variation across grade levels. As they note, this is not necessarily surprising given that social-emotional development starts in early childhood and is heavily influenced by out-of-school factors. Schools' relative influence on SE skills within CORE could also change over time as relevant data are incorporated into the SQIS. Yet these data do call into question efforts to use the measures to draw fine-grained distinctions in school performance. Using a Hierarchical Linear Model, for example, Hough et al. (2017) find that only 50% of CORE schools have estimated effects on a summary measure of SE skills that can be statistically distinguished from the collaborative-wide average.

In addition to providing evidence on the SE measures' internal reliability, Table 2 also reveals important differences in mean SE scores across student subgroups. In particular, English language learners and students with disabilities tend to rate their skills at the lowest levels across all of the SE measures, while white students consistently rate themselves most favorably. All of these differences are statistically significant ($p < 0.01$), and many are substantial in magnitude. For example, the self-management scores of students with disabilities lag those of the full sample by 0.41 standard deviations. In ongoing work, we are testing for measurement invariance across student subgroups in order to better understand the sources of these group differences and guide their interpretation. However, they clearly suggest the importance of taking into account student background characteristics when using these types of SE measures as an indicator of school performance. To this end, CORE's SQII incorporates and assigns weight not just to the overall SE score for each school but also to the scores of four subgroups: the lowest-performing racial or ethnic group, English language learners, students with disabilities, and students eligible for a free or reduced-price lunch. The extent to which this approach enables fair comparisons of schools' success in supporting the development of student SE skills despite differences in background characteristics will be a key area of analysis as the SQIS is implemented.

4.3. Validity

Given CORE's intention to use the SE measures as indicators of school performance, we begin our validity analyses of the field test data by examining the school-level correlations between each of the measures and multiple indicators of students' academic performance and behavior. We next examine the relationship between the SE measures and academic and behavioral indicators at the student level, comparing overall and within-school correlations to test for the presence of reference bias due to differential item functioning across schools. Finally, we provide evidence on the measures' correlation with teacher reports at the student level for the subsample of students for whom they are available.

Table 3 shows the school-level correlations between each measure and various outcome metrics separately for elementary, middle, and high schools; each school observation is weighted by the number of students for whom survey data is available. Of course, CORE's decision to incorporate SE measures into the SQII reflected a view that they capture aspects of student success that are not reflected in traditional academic indicators. One would therefore not expect them to be perfectly correlated with academic performance and behavior. To the extent that SE skills contribute to academic success, however, they should be positively related.

Overall these analyses indicate strong, statistically significant correlations between students' SE skills and concurrent academic and behavioral outcomes when aggregated to the school level. The first column shows the bivariate relationship between district-standardized GPA and the four SE measures. For elementary students, each of the SE measures is positively correlated with students' course grades, with self-management and self-efficacy showing the strongest relationships. We see a similar pattern in middle schools, but the relationships for growth mindset and social awareness are stronger than for elementary schools. At the high school level, a different set of social-emotional skills is most predictive of average GPAs: the strongest relationship is for growth

Table 3
School-level correlations of social-emotional measures with academic and behavioral indicators.

	GPA	ELA test score	Math test score	Percent suspended	Absence rate
Elementary schools					
Self-management	0.62	0.82	0.79	− 0.35	0.02
Growth mindset	0.38	0.67	0.65	− 0.13	− 0.28
Self-efficacy	0.61	0.60	0.56	− 0.17	0.07
Social awareness	0.33	0.58	0.53	− 0.19	0.03
N (schools)	67	586	586	102	156
Middle schools					
Self-management	0.65	0.85	0.85	− 0.29	− 0.54
Growth mindset	0.56	0.76	0.72	*− 0.18*	− 0.41
Self-efficacy	0.60	0.67	0.65	− 0.13	− 0.32
Social awareness	0.55	0.74	0.71	− 0.31	− 0.55
N (schools)	219	280	280	175	114
High schools					
Self-management	0.41	0.52	0.49	− 0.14	*− 0.36*
Growth mindset	0.49	0.47	0.39	− 0.15	− 0.32
Self-efficacy	0.26	0.12	*0.18*	0.03	− 0.20
Social awareness	0.44	0.74	0.74	− 0.07	− 0.50
N (schools)	184	120	120	153	35

Note. Each cell reports the school-level correlation between the SE measure and the outcome at the top of the column; school observations are weighted by the number of SE survey respondents. GPA is either for fall courses or, where available, the full year; among students with both fall and cumulative GPA, their correlation is 0.88. Percent suspended is the percentage of respondents receiving a suspension during the 2014–15 school year. Absence rate is the average percentage of days absent among respondents during the 2014–15 school year. All correlations are statistically significant at $p < 0.01$ except those in italics ($0.01 < p < 0.05$) and those underlined ($p > 0.05$).

mindset, followed by social awareness. While self-efficacy is strongly associated with GPA in elementary and middle school, the relationship in high school is weaker. In addition to establishing the predictive validity of the SE measures in the aggregate, these results therefore suggest intriguing patterns with respect to which SE constructs are most important for students' academic success across grade spans.

The next two columns of Table 3 confirm that the SE measures are also strongly correlated with students' test score performance in ELA and math. Among elementary and middle schools, self-management and growth mindset are the strongest predictors. Among high schools, where test scores are only available for grade 11, social awareness and self-management have the strongest relationship with test score performance.

The final two columns show the relationship between the SE measures and two indicators of student behavior: district-standardized measures of the share of students in a school who were suspended during the 2014–15 school year and average absence rates. With the exception of self-efficacy in high school, we see a negative relationship between each of the SE measures and the share of students suspended during the academic year. Not all of these correlations are statistically significant at conventional levels, however. In middle schools and high school, we also see consistent negative relationships between the SE measures and absences. These relationships are not evident in elementary school (except for growth mindset), where students may exercise less control over their attendance.

While Table 3 confirms that the SE measures included in the SQII are positively correlated with indicators of academic performance and behavior, these relationships are not exact. Students in some schools that are high-performing academically clearly report lower than expected SE skills, and vice versa. On one hand, this could reflect authentic variation in performance across academic and SE domains. On the other, it could be that students rate their SE skills more critically in some schools than in others, perhaps due to variation in normative expectations across schools. Our final set of analyses considered this possibility.

To examine the potential of reference bias to affect the comparability of student responses across schools, we compared the strength of the student-level relationships between each SE measure and academic performance indicators in the sample as a whole and within particular schools. Specifically, we regressed the indicators on each SE measure without (i.e., the "overall" estimate) and with (e.g., the "within-school" estimate) school fixed effects included in the estimation model. Table 4 reports these regression coefficients and the difference between them, along with tests of statistical significance.

If reference bias stemming from differences in normative standards across schools were a significant concern, we would expect the within-school correlations to be stronger than the overall correlations, as the latter would be biased downward. The results in Table 4, however, reveal the opposite pattern: the SE measures are more strongly related to academic outcomes across the sample as a whole than when only

Table 4
Student-level correlations between social-emotional skills and academic indicators, overall and within-school.

	GPA			ELA test score			Math test score (Math)		
	Overall	Within	Diff	Overall	Within	Diff	Overall	Within	Diff
Elementary schools									
Self-management	0.44	0.41	0.03 +	0.37	0.3	0.07***	0.34	0.27	0.07***
Growth mindset	0.25	0.23	0.02	0.3	0.24	0.06***	0.28	0.22	0.06***
Self-efficacy	0.43	0.41	0.02*	0.29	0.25	0.04***	0.3	0.26	0.04***
Social Awareness	0.28	0.27	0.01	0.22	0.17	0.05***	0.18	0.14	0.04***
N	7893			111,902			111,902		
Middle schools									
Self-management	0.38	0.35	0.03***	0.35	0.28	0.07***	0.34	0.27	0.07***
Growth mindset	0.22	0.19	0.03***	0.36	0.31	0.05***	0.34	0.29	0.05***
Self-efficacy	0.33	0.31	0.02***	0.28	0.23	0.05***	0.31	0.27	0.04***
Social awareness	0.23	0.21	0.02***	0.2	0.15	0.05***	0.19	0.14	0.05***
N	116,645			110,293			110,293		
High schools									
Self-management	0.29	0.28	0.01***	0.21	0.19	0.02***	0.18	0.16	0.02**
Growth mindset	0.19	0.17	0.02***	0.28	0.26	0.02 +	0.22	0.21	0.01
Self-efficacy	0.27	0.27	0 +	0.14	0.15	− 0.01	0.19	0.2	− 0.01
Social awareness	0.18	0.17	0.01***	0.18	0.14	0.04***	0.15	0.11	0.04***
N	127,134			24,294			24,294		

Note. "Overall" columns report bivariate student-level correlations between each SE measure and the relevant outcome. "Within" columns report the same correlation after adjusting for school fixed effects. "Diff" column reports the difference between the overall and within correlations. Standard errors (not reported) are clustered by school. All overall and within correlations are statistically significant at $p < 0.001$. For differences, ***$p < 0.001$, **$p < 0.01$, *$p < 0.05$, +$p < 0.10$.

Table 5
Student-level correlations between self- and teacher reports of student social-emotional skills.

	Self-Mgt (student)	Soc-Awr (student)	Self-Mgt (teacher)	Soc-Awr (teacher)
Elementary schools [N = 8028]				
Self-Mgt (student)	1.00			
Soc-Awr (student)	0.52	1.00		
Self-Mgt (teacher)	0.41	0.23	1.00	
Soc-Awr (teacher)	0.36	0.22	0.85	1.00
Middle schools [N = 11,290]				
Self-Mgt (student)	1.00			
Soc-Awr (student)	0.51	1.00		
Self-Mgt (teacher)	0.38	0.23	1.00	
Soc-Awr (teacher)	0.32	0.21	0.82	1.00
High schools [N = 12,510]				
Self-Mgt (student)	1.00			
Soc-Awr (student)	0.55	1.00		
Self-Mgt (teacher)	0.47	0.28	1.00	
Soc-Awr (teacher)	0.42	0.26	0.85	1.00

Note. Table reports Pearson correlation coefficients. For 9376 students who were rated by more than one teacher, teacher-reports are averaged across 2–4 teachers. Self-Mgt: self-management; Soc-Awr: social awareness. All correlations are statistically significant at $p < 0.0001$.

comparing students attending the same schools. This analysis does not rule out the possibility that reference bias may lead to misleading inferences about specific schools. It does, however, provide preliminary evidence that the form of reference bias that would be most problematic for using SE data to evaluate school performance is not an important phenomenon across the CORE Districts as a whole.

As a final validity check on students' ratings of their SE skills, Table 5 reports student-level correlations between student self-reports and teacher reports within the two districts that administered teacher surveys during the field test. Teacher reports on the constructs of self-management and social awareness are available for approximately 31,828 students, who we examine separately by grade span. For self-management, student and teacher reports are modestly correlated, ranging from 0.38 in middle school to 0.47 in high school (all statistically significant at $p < 0.001$). The parallel correlations for social awareness, although also highly statistically significant, are weaker, at 0.21 in middle school and 0.26 in high school, consistent with the notion that this construct is less externally observable. Table 5 also shows that teachers' ratings of the same student's self-management skills and social awareness are correlated at 0.82 or higher for all grade levels, suggesting that many teachers may evaluate students holistically rather than draw distinctions between different constructs.

While some CORE Districts continue to collect teacher reports of students' SE competencies in addition to student self-ratings, teacher ratings are not currently included in the SQII. In some districts, teacher ratings of students' SE competencies were subject to the collective bargaining process and therefore could not be implemented unilaterally; other districts viewed evaluating students' competencies as part of educators' everyday responsibilities and preferred not to ask them to complete a separate assessment.

5. Building capacity and supporting school practice change

While refining its approach to measuring SE skills, CORE instituted a variety of mechanisms that draw on the data included in the SQII to support districts in changing school practice and building educators' capacity in the domain of social-emotional learning. These efforts include an online platform to share Index results with schools and various professional learning opportunities and resources for educators. CORE also designed a system of more intensive capacity building structures for two groups of schools identified as needing improvement. Below we describe these structures and supports, the early implementation of which is documented in detail in Marsh et al. (2017).

5.1. Index reports

CORE has developed comprehensive Index Reports for each school designed to be user-friendly for those without expertise in data analysis. The report enables a user to view all of the SQII indicators, including two-year trends and comparisons with other schools and districts. The fall 2016 reports (based on data from the 2015–2016 school year) include measures of achievement growth in English language arts and math based on statewide tests, as well as results from the SE and CC surveys.

5.2. Survey reports

Several districts within CORE additionally receive more detailed summaries of the survey-based SE and CC data from their survey administration contractor, Panorama Education, while other districts self-administer the surveys and create their own reports. Staff members in all CORE Districts are able to disaggregate their data by item, competency, school, and subgroup. This enables each district to determine which schools may need additional supports and identify opportunities to eliminate gaps between subgroups.

5.3. Intensive school support and capacity building

CORE's SQIS was inspired by the work of Fullan (2011), whose research emphasizes the importance of educators' intrinsic motivation to help all students succeed. As such, the SQIS focuses on providing supports to build schools' capacity rather than attaching punitive consequences to Index results. All schools and districts receive support from CORE in the form of professional learning resources such as interim assessment tools for English language arts and Math and peer learning opportunities related to school districts' areas of interest.

5.4. Communities of practice

A set of “focus” schools identified based on low academic performance by one or more student subgroups or large achievement gaps were grouped into Communities of Practice. These groups typically comprise between two and four schools within the same district and serve as an opportunity for a community of educators to focus on a common challenge using the Plan-Do-Study-Act (PDSA) model of rapid-cycle continuous improvement.

- *Plan*: Define the “problem of practice” or issue the group is attempting to address. Learn about ways to address the problem of

practice, determine which intervention or strategy to test, and decide which data will be used to determine the efficacy of the intervention
- *Do*: Implement the intervention or strategy
- *Study*: Collect and examine evidence about the efficacy of the intervention or strategy.
- *Act*: Based on that evidence, decide on the next steps: e.g., repeat the intervention, and explore a different intervention.

Each focus school develops a two-year action plan based on its own needs assessment and student data analysis and engages in the PDSA cycle three times per year. Participants in the Communities of Practice document teaching approaches, interventions attempted, and trainings provided for teachers and staff. They consider this information alongside SQII results and other school data to analyze which actions supported school improvement and which did not. Key lessons are then reported to each school's Site Council, which considers this information in conjunction with the school's state-required school improvement plan. While the Site Council is involved in the process, receiving updates on the progress of each Community of Practice, the district is responsible for determining whether the Community of Practice is helping to improve the school's outcomes in specific areas of need.

CORE's role with respect to the Communities of Practice is to provide data, tools, and resources to inform the planning efforts and professional development for facilitators. Although it is too early to determine the effectiveness of the Communities of Practice model in improving student outcomes, a CORE survey of community of practice participants from the 2014–15 school year indicated that approximately 75% of respondents agreed or strongly agreed that the program helped their school improve.

5.5. School pairings

Schools identified as “priority” schools due to low overall Index performance or graduation rates are paired with higher-performing “reward” schools that serve demographically similar students, which share best practices and provide technical assistance. CORE intends for these School Pairings to foster peer accountability and to be valuable to both participating schools. Data from a 2014–2015 survey of roughly 50 school pairing participants again supports this notion: nearly 80% of “reward” schools and 70% of “focus” or “priority” schools agreed or strongly agreed that participating in the program helped their school to improve.

Paired schools engage in peer learning by focusing on one or more SQII metrics that align with the priority school's needs assessment. Low-performing schools that are receiving funds through a federal School Improvement Grant (SIG) continue their work to implement the school improvement plan they outlined during the SIG process. Schools without an existing school improvement plan develop an action plan based on turnaround principles articulated in the CORE waiver application, and their pair schools support plan development and implementation through strategies such as school visits, joint PSDA cycles, coaching, and professional learning communities.

CORE's role in supporting school pairings has been to identify schools for intervention and support and to use Index data and input from district leaders to match them with a partner school. While each district leads its own improvement process, CORE provides administrative support and professional development through Pairing Institutes focused on building relationships between schools, developing initial plans for the pairings, and sharing continuous improvement strategies. CORE has thus far paired schools based solely on achievement and graduation data because the SE and CC data were not included in the SQII prior to the 2015–16 school year. However, paired schools have used SE and CC data from the field test to inform their improvement efforts.

6. Areas for continued exploration

The CORE Districts' waiver expired in August 2016, having been rendered moot by the enactment of the Every Student Succeeds Act. The districts have continued to implement the SQIS on a voluntary basis, however, and CORE has opened its data-sharing collaborative to other districts and charter school networks in California. By participating in the data-sharing collaborative, districts gain access to the full range of indicators and benchmarks included in the SQII. Participating districts will also participate in biannual convenings to discuss common challenges and share lessons. CORE has also begun to implement several new initiatives related to the SQIS and the development of students' SE competencies. These next steps, outlined below, may be instructive for other states or districts that choose to build upon CORE's work.

6.1. Evaluating SEL practices and the SQIS

CORE's 1500 schools employ a wide range of interventions, instructional approaches, and curricula to build students' SE skills. For example, individual CORE Districts have:

- Provided professional development for educators using growth-mindset and self-management toolkits developed by TransformEd, as well as a variety of other resources.
- Created a developmentally appropriate curricular scope and sequence for improving academic habits, motivation, well-being, and school CC.
- Developed a rubric for identifying observable markers of effective instructional practices to build SE skills.
- Mapped intersections between SE skills and the Common Core State Standards to illustrate how academics and SEL can be integrated.
- Added teacher ratings of students' SE skills to student report cards to ensure that students, teachers, and parents are having regular conversations about students' SE development.

CORE also has entered into a partnership with Policy Analysis for California Education (PACE), a research center based at Stanford University, the University of Southern California, and the University of California—Davis that will support member districts in identifying promising practices to support students' SE development. Using the SE measures included in the SQII, PACE will identify a sample of schools that demonstrate particularly strong SE outcomes for students. The PACE team will then conduct site visits and interviews to develop hypotheses about the practices that may be driving these outcomes. This information and CORE's peer-learning infrastructure will enable the districts to more rigorously evaluate the effectiveness of these practices and bring the most effective ones to scale, potentially offering a range of options based on students' particular SE strengths and needs. Through its partnership with PACE, CORE also aims to develop the relationships, data systems, and infrastructure to support a broader research agenda that includes evaluating the impact of key components of CORE's accountability model, such as Communities of Practice and School Pairings, and comparing the CORE model to other accountability frameworks in California.

6.2. Supporting cross-district learning around SEL

To foster more direct connections with teachers and district staff, TransformEd and the CORE Districts are piloting a year-long Social-Emotional Learning Fellowship for district staff members. SEL Fellows remain full-time district employees while also playing a leadership role in shaping SEL-related work across the CORE Districts. Each fellow gathers input from educators in his or her district to refine CORE's SEL-related survey administration, data reporting, and practice improvement work. For example, Fellows identify needs in their own districts, help plan and facilitate CORE-wide trainings related to SEL, and

develop tools that can support their counterparts in other CORE Districts.

To inform this work, TransformEd and CORE recently conducted semi-structured interviews with district leaders, principals, and teachers aimed at understanding how educators are interpreting and using the SEL data included in the SQII and where they perceive the need for more support. Although this work will be refined through further interviews and focus groups, initial analyses suggest the following opportunities for cross-district learning:

6.2.1. Building collective ownership

Because research on SEL is complex, it can be challenging to communicate succinctly what SE skills are and why they matter. One district administrator said, "District folks don't necessarily have a firm… understanding of SEL, and it is hard for us to communicate…how transformative [SEL] is for students." Staff members from multiple CORE Districts have identified a need to develop more coherent messaging in order to articulate how SEL relates to other district priorities and to establish a collective sense of ownership for it.

6.2.2. Integrating SEL and academics

District staff members with expertise in SEL believe that SE skills must be fully integrated into academic content and instruction in order to improve student outcomes. One interviewee, for example, suggested that support for SEL "must be integrated with professional development on instructional strategies in academic content." While standalone SEL programs can be helpful resources, they sometimes convey that SEL is separate from schools' academic work. At the district level, the false division of "academics" and "student supports" (which often exist as separate departments) often creates a structural barrier to integrating these two interrelated areas.

6.2.3. Connecting data to instruction

Interviewees felt that CORE had done a great job of sharing information about the SQII, "especially the 40% that's non-academic," but indicated that more training and support could help connect the data to research and instructional strategies. For example, one principal said, "I have been using SE data from the district in our staff meetings to discuss what interventions would be helpful for students and to plan tier 1 and tier 2 interventions." Annual assessments and school-level data will likely be insufficient to improve practice, however. A supplemental, formative approach to measurement could include assessing individual students' SE skills regularly throughout the school year in a way that reinforces students who demonstrate growth and provides more granular data for teachers to use in refining their practice.

7. Lessons for the field

CORE's experience to date incorporating SE measures into its system of school accountability and continuous improvement suggests several lessons that can inform other states' and districts' efforts to use innovative measures at scale as part of an expanded definition of student success.

7.1. Articulate key principles and non-negotiables up front

For CORE, these principles were college- and career-ready expectations for all students; a focus on collective responsibility rather than high-stakes accountability; the development of intrinsic motivation for change through differentiated recognition, accountability, and support for schools; and focused capacity-building for effective instruction and leadership, especially in schools identified as needing improvement. These principles were used to inform decisions about implementation, guide the development of workarounds and solutions as challenges arose, and clearly signal the system's priorities to stakeholders.

7.2. Ensure district buy-in

The endorsement of the SQIS by the individual districts participating in CORE has been critical to its successful implementation. District buy-in stemmed from the fact that district leaders co-designed and opted into an approach to accountability and continuous improvement that reflected their beliefs and values. They also made joint decisions about performance indicators and system implementation through participation on CORE's governing board and repeatedly reaffirmed their commitment to the approach by approving waiver amendments for submission to U.S. Department of Education. District staff had numerous opportunities to share feedback during the waiver implementation and to collaborate with each other in ways that supported professional growth and commitment to the SQIS.

7.3. Plan to iterate

No system of accountability and continuous improvement is perfect at the outset. When designing complex systems, it is important to be clear that the model will be refined over time based on findings that emerge from the data, feedback from stakeholders, and developments in research. Throughout the process of iterating, the key principles of the system must remain constant, providing educators and community members a sense of coherence. This was evident in CORE's waiver amendments: changes were requested to individual components of the waiver (e.g., to the method of calculating different indicators), but CORE's key principles remained consistent.

7.4. Roll out novel measurement approaches in stages

CORE tested and rolled out its new accountability system in several phases over three years, making many changes along the way based on both data and district feedback. This process increased buy-in by giving the districts time to understand how each measure works and to participate in field testing before new measures were incorporated into the Index. Additionally, the phased rollout process allowed time for district staff to build cross-district relationships, connect with others who were wrestling with similar challenges, and develop new approaches for acting on the data once the system was fully implemented.

7.5. Draw on support from external partners

In implementing the SQIS, CORE benefited from technical and financial support from a large number of nonprofit organizations, researchers, and funders. These partners helped with a range of tasks: supporting the selection of SE and CC competencies, validating the SE measures, analyzing the SQII data, developing data reports for schools, and providing professional development to participating districts. Ultimately, CORE was able to harness the skills and perspectives of many partners to build a robust system that no single district could have built alone (Knudson & Garibaldi, 2015).

8. Conclusion

There is widespread interest among education leaders and policymakers in the CORE Districts' School Quality Improvement System as an example of how to think more expansively about the definition of student success and the factors that support it, something ESSA enables all states to do. At the same time, some observers are understandably skeptical about the use of survey-based measures of social-emotional skills for accountability purposes (Duckworth & Yeager, 2015). Concerns regarding reference bias, incentives to "game" survey responses once stakes are attached, and the ability to differentiate schools' performance are well-founded. Continued research is needed to shed light on the empirical relevance of these concerns in the context of real-world school systems, including the districts participating in CORE.

Several factors may mitigate concerns about the use of survey-based SE measures within the context of CORE's accountability and continuous improvement system. First, CORE has so far assigned relatively little weight to the SE measures, with students' survey responses accounting for just 8 points of the 100-point School Quality Improvement Index. Second, the sole consequence for schools that perform poorly on the overall Index is to be paired with higher-performing schools that provide mentorship and support, marking a stark contrast with accountability systems that impose punitive sanctions on schools identified for improvement. Third, data from CORE's pilot and field test reveal positive correlations between the SE measures and students' academic and behavioral outcomes and no evidence to suggest that inter-school comparisons are undermined by reference bias. Although these data were gathered before the SE measures were formally included in the Index, these patterns suggest that the measures may be useful in guiding improvement efforts.

As states and districts consider using similar measures for formative or summative purposes, it is critical that they field test survey instruments, collect feedback from stakeholders as they are implemented, and stay abreast of the latest research in order to refine both measures and administration protocols. To avoid unintended consequences, careful attention should be paid to how much weight the measures are given in evaluating school performance, how the data are reported and used, and any stakes attached to results.

Ultimately, however, the use of student surveys of SE skills in systems of accountability and continuous improvement like CORE's presents an important learning opportunity for the field of social-emotional learning. It creates a context in which to study the properties of SE measures when implemented at scale and whether and how those properties change with repeated administration and when results are publicly reported. It creates infrastructure with which to better understand the development of students' SE skills over time and the effectiveness of school-based strategies to promote them. Perhaps most importantly, it provides a setting in which to identify the supports needed to ensure that educators respond to the provision of data on SE skills in ways that help more students succeed in school and life.

References

Almlund, M., Duckworth, A. L., Heckman, J. J., & Kautz, T. D. (2011). Personality psychology and economics. In E. A. Hanushek, S. Machin, & L. Woessmann (Vol. Eds.), *Handbook of the economics of education. 4. Handbook of the economics of education* (pp. 1–181). Amsterdam: Elsevier, North-Holland.

Bandura, A. (1997). *Self-efficacy: The exercise of control*. New York, NY: W.H. Freeman.

Collaborative for Academic, Social, and Emotional Learning [CASEL]. (2005). *Safe and sound: An educational leader's guide to evidence-based social and emotional learning programs—Illinois edition*. Chicago, IL: Author.

Duckworth, A. L., & Yeager, D. S. (2015). Measurement matters: Assessing personal qualities other than cognitive ability for educational purposes. *Educational Researcher*, *44*(4), 237–251.

Dweck, C. (2006). *Mindset: The new psychology of success*. New York, NY: Random House Publishing Group.

Feldt, L. S. (1969). A test of the hypothesis that Cronbach's alpha or Kuder-Richardson coefficient twenty is the same for two tests. *Psychometrica*, *34*(3), 363–373.

Fisher, R. J. (1993). Social desirability bias and the validity of indirect questioning. *Journal of Consumer Research*, *20*, 303–315.

Fullan, M. (2011). *Change leader: Learning to do what matters most*. San Francisco, CA: Jossey-Bass.

Goldman, S. (2006). Self-discipline predicts academic performance among low-achieving adolescents. *Res: A Journal of Undergraduate Research*, *2*(1), 84–97.

Heckman, J. J., Stixrud, J., & Urzua, S. (2006). The effects of cognitive and noncognitive abilities on labor market outcomes and social behavior. *Journal of Labor Economics*, *24*(3), 411–482.

Heine, S. J., Lehman, D. R., Peng, K., & Greenholtz, J. (2002). What's wrong with cross-cultural comparisons of subjective Likert scales? The reference-group effect. *Journal of Personality and Social Psychology*, *82*(6), 903–918.

Hough, H., Kalogrides, D., & Loeb, S. (2017). Using surveys of students' social-emotional learning and school climate for accountability and continuous improvement. *Policy analysis for California education*. Retrieved from http://www.edpolicyinca.org/publications/using-sel-and-cc.

Kendziora, K., & Osher, D. (2016). Promoting children's and adolescents' social and emotional development: District adaptations of a theory of change. *Journal of Clinical and Adolescent Psychology*, *45*(6), 797–811.

King, G., Murray, C. J. L., Salomon, J. A., & Tandon, A. (2004). Enhancing the validity and cross-cultural comparability of measurement in survey research. *American Political Science Review*, *98*(1), 191–205.

Knudson, J., & Garibaldi, M. (2015). *None of us are as good as all of us: Early lessons from the CORE districts*. American Institutes for Research. Retrieved from http://www.air.org/resource/none-us-are-good-all-us-early-lessons-core-districts.

Kyllonen, P. C., & Bertling, J. (2013). Innovative questionnaire assessment methods to increase cross-country comparability. In L. Rutkowski, M. von Davier, & D. Rutkowski (Eds.), *A handbook of international large-scale assessment data analysis: Background, technical issues, and methods of data analysis* (pp. 277–286). London, UK: Chapman Hall/CRC Press.

Marsh, J. A., Bush-Mecenas, S., & Hough, H. (2017). Learning from early adopters in the new accountability era: Insights from California's CORE waiver districts. *Educational Administration Quarterly*. http://dx.doi.org/10.1177/0013161X16688064.

Moffitt, T. E., Arseneault, L., Belsky, D., Dickson, N., Hancox, R. J., Harrington, H., ... Caspi, A. (2011). A gradient of childhood self-control predicts health, wealth, and public safety. *Proceedings of the National Academy of Sciences*, *108*(7), 2693–2698.

National Research Council (2012). Education for life and work: Developing transferable knowledge and skills in the 21st century. In J. W. Pellegrino, & M. L. Hilton (Eds.), *Board on Testing and Assessment and Board on Science Education, Division of Behavioral and Social Sciences and Education*. Washington, DC: The National Academies Press.

Podsakoff, P. M., MacKenzie, S., & Lee, J. (2003). Common method biases in behavioral research: A critical review of the literature and recommended remedies. *Journal of Applied Psychology*, *88*(5), 879–903.

Spencer, S. J., Steele, C. M., & Quinn, D. Q. (1999). Stereotype threat and women's math performance. *Journal of Experimental Social Psychology*, *35*(1), 4–28.

Walton, G. M., & Spencer, S. J. (2009). Latent ability grades and test scores systematically underestimate the intellectual ability of negatively stereotyped students. *Psychological Science*, *20*(9), 1132–1139.

West, M. R. (2016). Should non-cognitive skills be included in school accountability systems? Preliminary evidence from California's CORE districts. *Evidence Speaks Reports*. *1*Brookings Institution (13).

West, M., Kraft, M. A., Finn, A. S., Martin, R., Duckworth, A. L., Gabrieli, C. F. O., & Gabrieli, J. D. E. (2016). Promise and paradox: Measuring students' non-cognitive skills and the impact of schooling. *Educational Evaluation and Policy Analysis*, *38*(1), 148–170.

West, M. R., Dow, A. W., & Buckley, K. H. (2017). Using anchoring vignettes to enhance self-reports of social-emotional skills: Evidence from California's CORE districts. *Paper presented at the Annual meeting of the National Council on Educational Measurement, San Antonio, TX*.