A Comprehensive Examination of the School District of Philadelphia's Kindergarten Classroom Engagement Scale (CES): Validation Report

Katherine Barghaus, PhD

John Fantuzzo, PhD

Benjamin Brumley, BA

Kristen Coe, MSW

Whitney LeBoeuf, PhD

NOVEMBER 2017

Acknowledgements

The Penn Child Research Center at the University of Pennsylvania would like to thank the School District of Philadelphia's Office of Early Childhood Education and the William Penn Foundation for their support of scientifically sound assessment to foster the development of students' competencies in kindergarten. This work was completed with funding from the William Penn Foundation. The opinions expressed in this report are those of the authors and do not necessarily reflect the views of the William Penn Foundation or the School District of Philadelphia.

Suggested Citation:
Barghaus, K., Fantuzzo, J., Brumley, B., Coe, K. & LeBoeuf, W. (2017). *A Comprehensive Examination of the School District of Philadelphia's Kindergarten Classroom Engagement Scale (CES): Validation Report*. Philadelphia, PA: Penn Child Research Center.

Table of Contents

Executive summary

Success in school depends not only on academic skills but also on the development of social-emotional competencies. An important aspect of social-emotional learning is classroom engagement, which includes both academic and social engagement. Academic engagement reflects a student's approach to participating in learning activities, for example by demonstrating consistent effort and working independently. Social engagement reflects a student's approach to participating in the classroom community, such as working cooperatively (Finn & Zimmer, 2012). These skills have been shown to be observable, teachable, and readily incorporated into classroom instruction and home routines, making them prime candidates for sustainable interventions (McKown, 2017). It is not surprising then that the Every Student Succeeds Act (2015) calls for a broader definition of student success that includes indicators such as classroom engagement. To answer this call, we need evidence-based assessments of student engagement skills that can be tied to specific strategies for teachers and families to support students' development of these skills.

The Penn Child Research Center and School District of Philadelphia (SDP) have partnered to foster the classroom engagement skills of students entering public school kindergarten. To do so, a three-phase, evidence-based, plan was developed. The first phase was to establish the scientific validity and reliability of the District's measure of classroom engagement currently used at-scale with all kindergarteners–the *Classroom Engagement Scale* (CES). This measure consists of 14 items and it appears on the kindergarten report card which is sent home quarterly. The second phase would use a validated CES to develop a robust home-school intervention to enhance the development of these skills. The final phase would evaluate and improve the effectiveness of this intervention to strengthen student engagement.

This project completed the first phase of this plan by rigorously evaluating the CES to determine if its items relate to one another in a way that aligns with Academic and Social Engagement dimensions, if these dimensions can be applied to all student groups across time, and if these dimensions are related to student outcomes. The research questions were:

1. Is there scientific support for the CES capturing two distinct dimensions of Academic Engagement and Social Engagement?
2. Do the dimensions of the CES operate consistently across student subgroups (i.e., sex, race/ethnicity, English Language Learners, students with disabilities, and free/reduced lunch recipients)?
3. Are the dimensions of the CES stable across report card marking periods providing an evidence base for progress monitoring?
4. To what degree does a student's score on CES dimensions reflect the student's level of functioning, and not extraneous information about the teacher or classroom?
5. Do the identified CES dimensions relate to other academic and non-academic outcomes measured concurrently in kindergarten?
6. To what extent do the CES dimensions predict future student outcomes?

a. Do the identified CES dimensions predict outcomes in third-grade?

b. Are the CES dimensions predictive of third-grade outcomes above and beyond other kindergarten outcome measures?

The key findings for each of the research questions investigated are summarized below.[1]

- Results indicated that the CES reliably measures two dimensions of classroom engagement—Academic Engagement and Social Engagement.
- Findings indicated that the two dimensions operated consistently across the student subgroups examined and therefore can be used with these groups.
- Results revealed that the CES dimensions are stable across the kindergarten year, from the first to the fourth quarter, which allows for monitoring progress over time.
- Approximately a quarter of the variation in student scores was associated with some non-student source of variance (e.g. teacher or classroom characteristics).
- Analyses revealed that scores on both dimensions were significantly related to kindergarten mathematics grades, absences, and suspensions, and early literacy skills. Academic Engagement explained more variance in mathematics grades, absences, and early literacy skills while, Social Engagement explained slightly more variance in suspensions.
- Academic Engagement and Social Engagement in kindergarten both significantly predicted academic and non-academic outcomes in third-grade with Academic Engagement explaining more variance in Pennsylvania System of School Assessment scores and Social Engagement explaining slightly more variance in number of days suspended.
- Academic Engagement and Social Engagement in kindergarten were found to be significant predictors of third-grade outcomes *above and beyond* other kindergarten outcome measures.

Evidence Supported Uses and Next Steps

This report describes the successful completion of the first phase of a plan to foster greater classroom engagement among kindergarteners in the Philadelphia School District. Collectively the findings support the use of the CES by kindergarten teachers to assess students' engagement skills at kindergarten entry and monitor their development throughout kindergarten. The CES could now serve as the basis for developing evidence-based supports for the teachers and families to foster students' classroom engagement skill development. These supports could then be evaluated to test and improve their effectiveness. Ultimately, the CES and these supports could be extended from pre-kindergarten through third-grade to create a valuable mechanism to measure, monitor, and foster students' development of important social-emotional skills across these critical early grades. By doing so, the CES would make visible to teachers, administrators, and families the degree to which our youngest students are connected to learning and to the social support systems of the classroom community.

[1] Source: Findings derived from data provided by The School District of Philadelphia.

Introduction

Success in school depends not only on academic skills but also on the development of social-emotional competencies. An important aspect of social-emotional learning is classroom engagement, which includes both academic and social engagement. *Academic engagement* reflects a student's approach to participating in learning activities for example by demonstrating consistent effort and working independently. *Social engagement* reflects a student's approach to participation in the classroom community, such as working cooperatively (Finn & Zimmer, 2012). Research shows that the ability to engage academically and socially is a strong predictor of educational success and well-being (Matthews, Kizzie, Rowley, & Cortina, 2010) and that developing these skills early helps ensure that students begin on positive developmental trajectories (Jones, Barnes, Bailey, & Doolittle, 2017).

Classroom engagement skills have been shown to be observable, teachable, and readily incorporated into classroom instruction and home routines, making these skills good candidates for sustainable interventions (McKown, 2017). As such, it is not surprising that the new federal Every Student Succeeds Act (2015) calls for a broader definition of student success including "non-academic" indicators, such as student engagement. To meet this call, we need evidence-based assessments of students' engagement skills. To make sure these assessments are useful, they must be tied to specific practical strategies that teachers and families can use to support the development of academic and social engagement skills.

For decades, the School District of Philadelphia (SDP), in partnership with the Penn Child Research Center (PCRC), has invested in the development and use of quality teacher-report measurement of classroom engagement skills incorporated in the report card to share with families and school administrators. In the 1990s, SDP implemented a performance assessment battery included in the report card to assess students' cognitive skills and abilities, motor skills, and social-emotional learning competencies, including classroom engagement, across the early primary grades (Fantuzzo et al., 2005). As part of the development of this battery, PCRC worked with teacher leaders across grades to craft items that identified observable and teachable classroom competencies supported by the educational research literature. The subset of items measuring classroom engagement was validated by using other pertinent measures specifically developed for use within a large, diverse, urban context, including the Learning Behaviors Scale (Stott, McDermott, Green, & Francis, 1988) and the Adjustment Scales for Children and Adolescents (McDermott, 1993). The battery was designed to be a face-valid teacher tool that could be implemented without much formal training. However, prior research by PCRC in SDP has found that a third to two-thirds of the variability in students' scores on other teacher report measures is unrelated to students' functioning and more about teacher and classroom differences (Waterman, McDermott, Fantuzzo, & Gadsden, 2012). This work underscores the importance of assessing teacher variance in any measurement work involving teacher report of students' competencies.

In the early 2000s, the cognitive and motor assessments were removed from the report card and replaced with other assessments. However, the classroom engagement items were

retained on the report card to capture these important skills as students start their public education. This version of classroom engagement items was validated for use with all first-, second-, and third-grade students in the District (Barghaus, Fantuzzo, LeBoeuf, Henderson, Li, & McDermott, 2017). This research revealed that these items captured two dimensions of engagement (social and academic). Currently, 14 classroom engagement items are included on only the kindergarten report card which is sent to families four times during the academic year to report on student progress. Teachers rate each of these 14 engagement behaviors on a three-point scale: *Improvement Needed*, *Satisfactory*, or *Outstanding*. Collectively the current items are referred to as the *Classroom Engagement Scale (CES).* However, this set of items has not yet been examined to determine if they are reliable and valid for use in kindergarten with all student groups across report card periods.

PCRC has continued to partner with SDP's Office of Early Childhood Education to consider how to foster greater classroom engagement for students in public school kindergarten using the CES. In order to do so, a three-phase, evidence-based plan was developed. The first phase is to examine the validity and reliability of CES. Once an evidence base is established, the second phase is to use the CES as a foundation to develop robust home-school interventions to enhance the development of these skills. The final phase would be to evaluate and improve the effectiveness of these home-school interventions and strengthen family engagement in supporting students' development of these important engagement skills.

Scientific Validation Study

The primary purpose of this project was to conduct a rigorous and comprehensive investigation of the validity and reliability of the CES while used at scale in the SDP Kindergarten Report Card. This investigation consisted of determining if the current set of items on the Kindergarten Report Card relate to one another in a way that align with academic and social engagement dimensions, if these dimensions can be applied equally to relevant student groups across time, and if these dimensions are related to other relevant outcomes collected by SDP.

Evidence of the dimensional structure of a measure (construct validity) indicates that its items operate in a manner that is aligned with the constructs it purports to measure. The specific research questions posed to investigate the dimensional structure of the CES were:

1. Is there scientific support for the CES capturing two distinct dimensions of academic engagement and social engagement?
2. Do the dimensions of the CES operate consistently across relevant subgroups of students (i.e., sex, race/ethnicity, English Language Learners, students with disabilities, and free/reduced lunch recipients)?
3. Are the dimensions of the CES stable across report card marking periods providing an evidence-base for progress monitoring?
4. To what degree does a student's score on CES dimensions reflect the student's level of functioning and not extraneous information about the teacher or classroom?

Evidence based on the relations of the CES to other relevant measures (concurrent and predictive validity) indicates the extent to which CES dimensions are related to existing assessments of similar and dissimilar constructs administered to the same students at approximately the same point in time (concurrent) or in the future (predictive). The specific research questions posed to investigate the CES's relations to other variables were:

5. Do the identified CES dimensions relate to other important academic and non-academic outcomes measured concurrently in kindergarten?

6. To what extent do the CES dimensions predict other relevant future outcomes?

    a. Do the identified CES dimensions predict academic and non-academic outcomes in third-grade?

    b. Are the CES dimensions predictive of third-grade academic and non-academic outcomes above and beyond other kindergarten outcome measures?

This technical report is organized sequentially to report findings from each of the study research questions. First, the two cohorts of SDP kindergarten students providing data used for this study are described and the measures and variables employed are defined. Next, the findings for the research questions pertaining to the dimensional structure of the CES (construct validity) and its relations with other variables are presented (concurrent and predictive validity). Finally, the findings and the uses of the CES are summarized and recommendations for next steps are made. More detailed descriptions of the methodologies employed are presented in Appendix A and supplementary tables and figures are presented in Appendix B.

CES Validation Study Sample

The present study examined report card data from two cohorts of students enrolled in the School District of Philadelphia (SDP).2 The first cohort was the primary analytic sample that was used to investigate all research questions. The second cohort was used to replicate the dimensional structure found with the first cohort and to extend the external validity analyses by employing additional measures of early literacy competencies only collected with this cohort.

The first cohort consisted of all students enrolled in kindergarten during the 2011-2012 school year who had complete fourth quarter CES data and an identifiable primary teacher, which was necessary to conduct multilevel analyses that can account for the grouping of students within classrooms (n = 11,734). This represented 98.8% of the students enrolled in kindergarten in the school district during the fourth-quarter. To be included in the Cohort 1 concurrent validity sample (n = 10,894), a student was required to have complete data for the concurrent mathematics, suspension, and attendance outcomes. Students in the Cohort 1 predictive validity sample (n = 7,546) were students who were in third grade in SDP three years later (the 2014-2015 school year) and had complete outcome data for reading, math, suspensions, and

[2] Source: Findings derived from data provided by The School District of Philadelphia.

attendance. Table 1 presents the student demographic characteristics for the full Cohort 1 sample, as well as the sub-samples used for the concurrent and predictive validity analyses. Overall, there were minimal demographic differences between the three Cohort 1 samples (see Appendix B Figure 1 for a visual depiction of the Cohort 1 samples).

The second cohort (Cohort 2) consisted of students in kindergarten during the 2014-2015 academic year with complete fourth-quarter CES scores, an identified teacher, and complete scores on the AIMSweb assessment of early literacy competencies (*n* = 9,055). This study used scores from four AIMSwebs subtests: Letter Naming Fluency, Letter Sound Fluency, and Phonemic Segmentation Fluency, Nonsense Word Fluency.[3]

Table 1. **Demographic Characteristics of the Full, Concurrent, and Predictive Samples**

	Full		Concurrent		Predictive	
Variable	*n*	(%)	*N*	(%)	*n*	(%)
Race/Ethnicity						
Black/African American	6,016	51%	5,529	51%	3,700	49%
Hispanic/Latino	2,358	20%	2,183	20%	1,655	22%
White	1,740	15%	1,659	15%	1,070	14%
Asian	796	7%	755	7%	607	8%
Multi Racial/Other	809	7%	755	7%	503	7%
American Indian/Alaskan	15	0.13%	13	0.12%	11	0.15%
Sex						
Female	5,759	49%	5,390	49%	3,833	51%
Male	5,975	51%	5,504	51%	3,713	49%
Special Needs	619	5%	570	5%	343	5%
Limited English Proficiency	1,340	11%	1,229	11%	964	13%
Free and Reduced Lunch	8,578	73%	8,041	74%	5,346	71%
Total	11,734		10,894		7,546	

Measures

Below are brief descriptions of the measures and variables used in this study. Data on these indicators came from existing school district administrative data records.

The Pennsylvania System of School Assessment (PSSA). The PSSA is the Commonwealth of Pennsylvania's state standardized test of student achievement and includes measures of English Language Arts and Mathematics knowledge and skills. The reliability and validity of the PSSA scaled scores has been well established (Data Recognition Corporation, 2015) and includes high internal consistency (r range .92 to .94) and validity evidence from factor analysis and differential item functioning.

[3] Cohort 2 is smaller than Cohort 1 for several reasons including: (1) kindergarten enrollment in the District dropped by 4% between the 2011-12 and the 2014-15 academic school year; and (2) the AIMSweb came into widespread use in Philadelphia in the spring of 2015, after a pilot period, however, not all kindergarten students complete the assessment and therefore they were not included in the analytic sample for Cohort 2 (18% of students with CES data and an identifiable teacher did not have AIMSweb scores).

Kindergarten Mathematics Grade. Kindergarten mathematics grade reflected fourth-quarter kindergarten mathematics achievement reported by teachers on a continuous scale from 0 to 100.

AIMSweb. The AIMSweb literacy assessment system consists for four tests of early literacy competencies: Letter Naming Fluency, Letter Sound Fluency, and Phonemic Segmentation Fluency, Nonsense Word Fluency. Test-retest reliability estimates of .81 and .82 are reported for Letter Naming Fluency and Letter Sound Fluency scores, respectively (Pearson, 2012). For Phonemic Segmentation Fluency and Nonsense Word Fuency scores average alternative-form reliability estimates of .61 and .74 are reported, respectively (Pearson, 2012). The developers also report criterion validity evidence for all four scores (Pearson, 2012).

Attendance. Attendance was calculated as the total number of absences (excused, unexcused, or out of school suspension) during the academic year.

Suspension. The number of suspension days was calculated for each student by totaling the number suspension days within a given academic year.

Sex. Student sex was indicated as either male (1) or female (0).

Race/ethnicity. For each student one of the following race/ethnicity categories was indicated: African American, White, Hispanic, Asian, American Indian/Alaskan Native, or multi-racial/other.

English Language Learners. School district enrollment records indicated whether a student was classified as "Limited English Proficient" (LEP) during a given school year. English proficiency was coded as 1 for LEP and 0 for non-LEP.

Special Education. Special education status was coded as 1 for students in special education and 0 for student who did not participate in special education.

Free or Reduced Price Lunch. Students qualifying for free or reduced lunch were coded as 1 and students who did not qualify or whose parents did not apply were coded as 0.

Findings: Dimensions of the CES

Below we outline the key findings for each research question and briefly describe the data analytic methods that were conducted to produce them. Additional details on the analyses employed can be found in Appendix A.

Research Question 1. Is there scientific support for the CES capturing two distinct dimensions of academic engagement and social engagement?

ANSWER: Yes. Analyses of data from Cohort 1 and Cohort 2 identified two reliable dimensions that best represent the CES data–*Academic Engagement* and *Social Engagement*. These dimensions are defined as:

1. *Academic Engagement*: reflects a student's approach to participating in learning activities and consists of 5 items that reflect behaviors such as attentiveness to completing academic tasks (e.g., "Strives for quality work") (Cronbach's alpha = .92).

2. *Social Engagement*: reflects a student's approach to participation in the classroom community and consists of 7 items that rate skills such as appropriately interacting with teachers and other students (e.g., "Works and plays cooperatively with others") (Cronbach's alpha = .95).

To determine the number and composition of distinct dimensions measured by the CES, multilevel exploratory factor analyses (EFA) and multilevel confirmatory factor analyses (CFA) were used. EFA is used to uncover the number of distinct dimensions that best describe the data based on how the indicators relate to one another. CFA is used to test the fit of hypothesized dimensions to the data. These analyses indicated that a two-dimensional structure best represented the data. Since a large inter-factor correlation (r = .81) was found, the presence of a general dimension was also investigated. However, this model, with one General Engagement factor measured by all 14 items and two specific factors derived from EFA, did not fit the data better than the two-dimensional structure (see Appendix B Table 2 for model fit statistics). Thus, the two-dimensional structure was retained for further testing.

***Table 2*. Classroom Engagement Scale Dimensions**	
Academic Engagement	**Social Engagement**
Completes work on time	Handles conflict appropriately
Can work independently	Respects others rights/diversity/feelings/property
Demonstrates consistent effort	Works and plays cooperatively with others
Strives for quality work	Accepts responsibility for choices and actions
Participates in group activities	Makes appropriate movement between activities
	Listens and follows directions
	Respects school environment and materials

Note. The interfactor correlation was .81. Items are presented in order of the magnitude of their factor loading with those items with the largest loading listed first. See Appendix B Table 1 for factor loadings.

Research Question 2. Do the dimensions of the CES operate consistently across relevant subgroups of students (i.e., sex, race/ethnicity, English Language Learners, students with disabilities, and free/reduced lunch recipients)?

ANSWER: Yes. Findings supported using the two-dimensional structure across *all* subgroups of students examined. This indicates that scores on the dimensions can be calculated for all students in these groups.

Multiple group confirmatory factor analysis, using robust estimation to account for the nesting of students within teachers, was used to examine the extent to which the CES dimensions are

equivalent across student groups. If the dimensions are equivalent across groups, the same model can be used to estimate scores for students in each group (Kim & Yoon, 2011). This is known as measurement invariance and to test it a series of models were estimated to test whether the CES operates consistently across groups (Millsap & Yun-Tein, 2004). These series of models are then compared to determine the extent to which measurement invariance holds. Comparisons were made using changes in model fit indices (Cheung & Rensvold, 2002; Chen, 2007). These analyses demonstrated minimal changes in the model fit indices indicating measurement invariance of the two-dimensional structures across sex, race/ethnicity, English Language Learners, students with disabilities, and free/reduced lunch recipient subgroups. Appendix B Table 3 contains the model fit statistics for all of these analyses.

Research Question 3. Are the dimensions of the CES stable across report card marking periods providing an evidence-base for progress monitoring?

ANSWER: Yes. Results supported using the two-dimensional structure across time. This indicates that the scores on the CES dimensions can be used to monitor students' development of these skills in kindergarten.

The multiple group confirmatory factor analysis was again used to examine the extent to which the CES dimensions are equivalent across time (Liu, Millsap, West, Tein, Tanaka, & Grimm, 2017). Specifically, we examined the extent to which the CES dimensions in the first report card marking period operated the same way in the fourth marking period to provide support for using scores on the dimensions to monitor progress (see Appendix B Table 3 for model fit statistics).

Research Question 4. To what degree does a student's score on CES dimensions reflect the student's level of functioning and not extraneous information about the teacher or classroom?

ANSWER: The findings indicate that 26% to 27% of the variance in the CES Academic Engagement and Social Engagement scores, respectively, was not directly attributable to the student but rather to some other sources of variance, such as the teacher or classroom.

Multilevel modeling was used to assess the extent to which the CES scores reflected information about individual students compared to other sources of variance, such as the teacher or the classroom context. The latter source of score variance can be problematic because it limits the ability of the measure to accurately differentiate children's true abilities. To investigate this, two-level HLMs were estimated for scores on each of the CES factors. Although scores on the CES provide information about children's ability, more than a fourth of the variability in scores may reflect something other than children's individual ability. Some non-student variation in scores is expected in teacher-report assessments. Similar, and in some cases higher, levels of non-student variation have been found for other teacher-report measures (see e.g., Waterman, McDermott, Fantuzzo, & Gadsden, 2012). Collectively, these findings point to the need to increases the amount of individual student information captured by teacher-report measures for example by providing additional assessment administration supports for teachers.

Findings: Relations of the CES to Other Variables

Research Question 5. Do the identified CES dimensions relate to other important academic and non-academic outcomes measured concurrently in kindergarten?

ANSWER: Yes. Analyses provided support for the concurrent validity of the Academic Engagement and Social Engagement dimensions of the CES. Analyses revealed that:

1. Academic Engagement and Social Engagement scores in kindergarten were significantly related to kindergarten mathematics grades, number of absences, number of days suspended, and AIMSweb scores.[4]

2. Academic Engagement scores in kindergarten explained more variance in kindergarten fourth quarter mathematics grades (31% versus 13%) and all four AIMSweb subtest scores (14% to 20% versus 5% to 8%) than Social Engagement scores. It also explained slightly more variance in the number of kindergarten absences than Social Engagement scores (3% versus 1%).

3. Social Engagement scores in kindergarten explained slightly more variance in the number of days suspended in kindergarten than Academic Engagement scores (4% versus 2%).

Concurrent validity is the extent to which a measure is related to other independent measures of similar and dissimilar constructs administered to the same students at approximately the same point in time. The relations between scores on the Academic Engagement and Social Engagement dimensions of the CES and other academic and non- academic outcomes measured in kindergarten were examined. Seven student outcomes were investigated: number of days absent, number of days suspended, fourth-quarter mathematics grade, and scores on four subtests of the AIMSweb (i.e., Letter Naming Fluency, Letter Sound Fluency, and Phonemic Segmentation Fluency, Nonsense Word Fluency). For all outcomes, two-level multilevel were estimated to partition the variance in the kindergarten outcomes into two components: (a) variance explained by non-student sources, such as the teacher assessor or classroom, and (b) variance related to differences in students' engagement behavior. For each outcome, the percentage of student-level variance that could be explained by the Academic and Social engagement scores was calculated (Selya, Rose, Dierker, Hedeker, & Mermelstein, 2012).

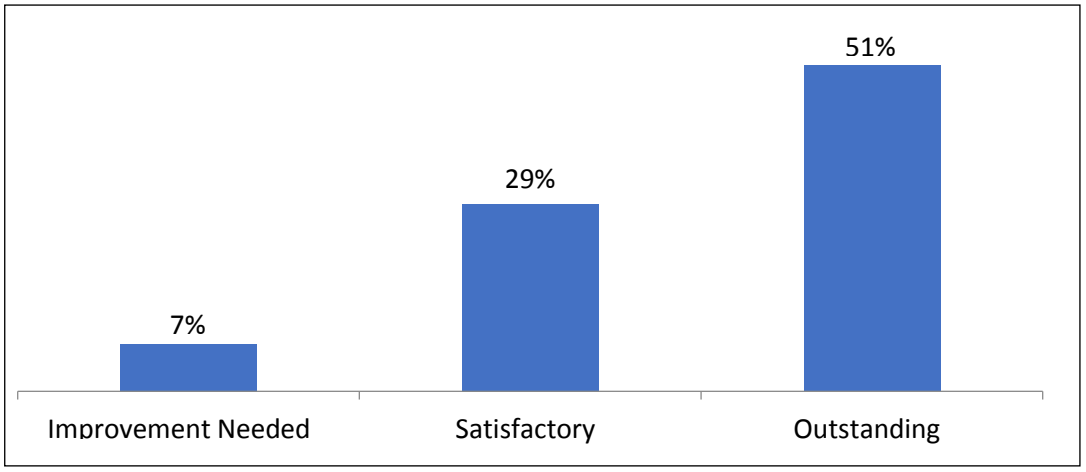
Research Question 6a. Do the identified CES dimensions predict academic and non-academic outcomes in third-grade?

ANSWER: Yes. Predictive validity analyses indicated that CES Academic Engagement and Social Engagement scores in kindergarten significantly predicted other relevant academic and non-academic outcomes in third-grade. Specifically, the findings revealed that:

[4] Suspensions were a low likelihood outcome with a positively skewed distribution in both kindergarten and third-grade. The result reported here are preliminary and should be interpreted with caution.

1. Academic Engagement and Social Engagement scores in kindergarten were significant predictors of third-grade PSSA scores, number of absences, and number of days suspended. For example, 51% of students rated "Outstanding" on Academic Engagement in kindergarten met proficiency on the third-grade English Language Arts PSSA, while only 7% of students rated as "Improvement Needed" met proficiency.

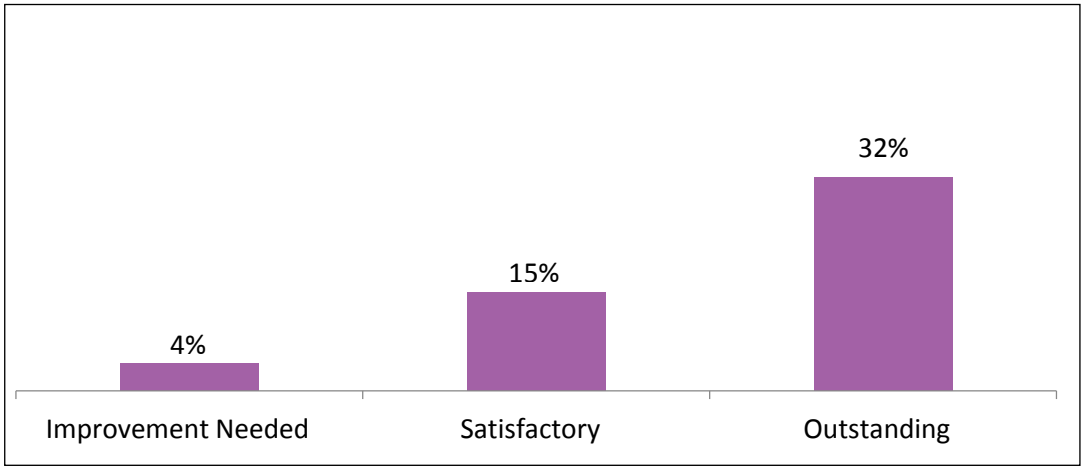
Figure 1. Proficiency rates on the third-grade PSSA English Language Arts test by average Academic Engagement score in kindergarten.



Note. Groups were created from average CES scores. Average scores of 1.5 to 2 were categorized as "Outstanding", 0.5 to 1.49 as "Satisfactory", and 0 to 0.49 as "Improvement Needed."

2. Academic Engagement scores in kindergarten explained more variance in third-grade English Language Arts and Mathematics PSSA scores than kindergarten Social Engagement scores (17% versus 9% and 15% versus 8%, respectively).

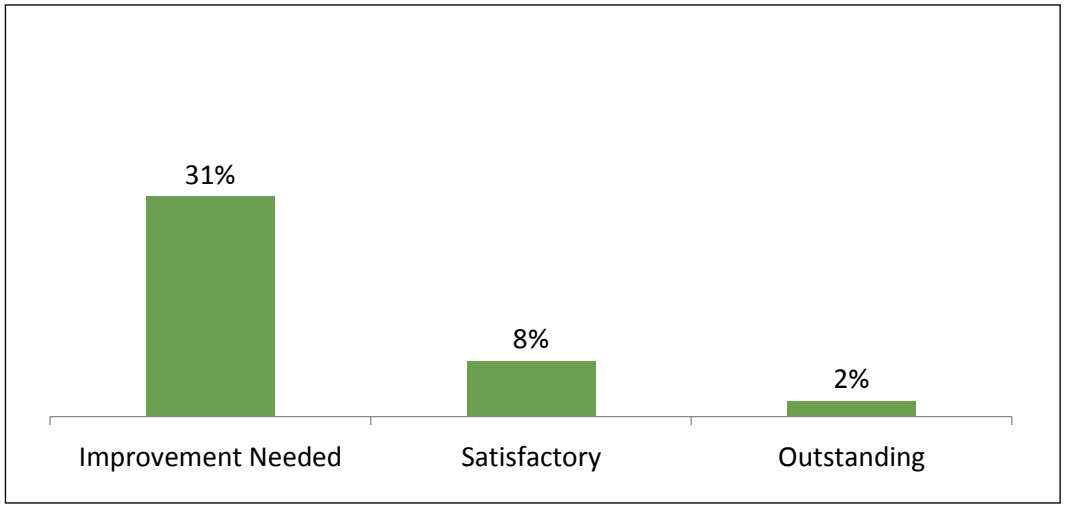
Figure 2. Proficiency rates on the third-grade PSSA Mathematics test by average Academic Engagement score in kindergarten



Note. Groups were created from average CES scores. Average scores of 1.5 to 2 were categorized as "Outstanding", 0.5 to 1.49 as "Satisfactory", and 0 to 0.49 as "Improvement Needed."

3. Social Engagement scores in kindergarten explained slightly more variance in the number of days suspended in third-grade than kindergarten Academic Engagement scores (5% versus 3%).

Figure 3. Suspension rate in third-grade by average CES Social Engagement score in kindergarten



Note. Groups were created from average CES scores. Average scores of 1.5 to 2 were categorized as "Outstanding", 0.5 to 1.49 as "Satisfactory", and 0 to 0.49 as "Improvement Needed."

Predictive validity is the extent to which one measure is related to other measures of similar and dissimilar constructs administered to the same students at different points in time. The relations between scores on the Academic Engagement and Social Engagement dimensions of the CES and the number of absences, and number of days suspended, and Mathematics and English Language Arts PSSA scores in third-grade were examined. For all outcomes, two-level multilevel were estimated to partition the variance in the third-grade outcomes into two components: (a) variance explained by non-student sources, such as the teacher assessor or classroom, and (b) variance related to differences in students' engagement behavior. For each outcome, the percentage of student-level variance that could be explained by the Academic and Social engagement scores was calculated (Selya, Rose, Dierker, Hedeker, & Mermelstein, 2012).

Research Question 6b. Are the CES dimensions predictive of third-grade academic and non-academic outcomes above and beyond other kindergarten outcome measures?

ANSWER: Yes and no. Scores on the Academic Engagement and Social Engagement dimensions in kindergarten were found to be significant predictors of third-grade outcomes *above and beyond* other kindergarten outcomes measures. This indicates that the CES scores provide information about students that *uniquely* predicts future outcomes. Specifically, these analyses indicated that:[5]

[5] Because f^2 values are scaled to reflect the proportion of variance explained relative to variance explained by the

1. Academic Engagement scores in kindergarten explained approximately an additional 9% of the variance in PSSA Mathematics scores above and beyond fourth-quarter kindergarten mathematics grades.

2. Academic Engagement scores in kindergarten explained approximately an additional 5% of the variance in PSSA English Language Arts scores above and beyond kindergarten reading level.

3. Social Engagement scores in kindergarten explained approximately an additional 4% of the variance in the number of days suspended in third-grade above and beyond days suspended in kindergarten.

4. Academic Engagement scores in kindergarten *did not* explain significantly more variance in the number of third-grade absences above and beyond the number of kindergarten absences.

To determine the unique predictive utility of the CES dimensions, a sequential series of models were run to test how much additional variance in third-grade outcomes could be explained by the CES dimension scores *above and beyond* scores on other kindergarten measures of the same or a similar outcome. Cohen's f^2, an effect size measure of variance explained in a multilevel regression model framework, was used to estimate the proportion of variance in third-grade outcomes uniquely accounted for by the CES, above and beyond the kindergarten measures (Seyla, Rose, Dierker, Hedeker, & Mermelstein, 2012). Using Cohen's (1992) criteria for small (values between .02 and.15), medium (between .15 and .35), and large (greater than .35) f^2 effect size, all of the findings presented above would be considered small effect sizes.

Evidence Supported Uses

Rigorous validity analyses provided support for two important classroom engagement dimensions of the CES from the educational research literature: Academic Engagement and Social Engagement. *Academic engagement* reflects a student's approach to participating in learning activities, for example by demonstrating consistent effort and working independently. *Social engagement* reflects a student's approach to participating in the classroom community, such as working cooperatively with teachers and peers. The evidence from this study supports the use of these dimensions for all student groups examined and supports their use across time to inform instruction and monitor progress. Most importantly, these dimensions were found to be related to other important academic and non-academic outcomes in kindergarten. The dimensions were also predictive of third-grade outcomes and provided unique information above and beyond other kindergarten predictors. Collectively this evidence supports the use of the CES by kindergarten teachers to assess student engagement skills at kindergarten entry and monitor their development throughout the kindergarten year.

---

full model they cannot be interpreted directly as a proportion of variance explained (Selya , Rose, Dierker, Hedeker, & Mermelstein, 2012). However, values closer to zero will closely match variance explained calculations. Given the magnitude of the findings reported here we provided this approximate variance explained interpretation.

Next Steps

Although the CES is used District-wide four times a year with thousands of kindergarteners, currently there are no resources available that explain what skills it measures, why they are important, or how to promote these skills in the classroom and at home. This lack of information and training impedes our ability to capitalize on the potential of the rich information provided by the CES. The evidence generated by this study provides a solid foundation for the second phase of work to bridge this gap–developing a sustainable system of home-school supports that foster student engagement skill development. This system should include two key components:

1. First, best classroom practices currently in place to support students' engagement skill development and evidence-based practices translated into specific instructional strategies and activities should be incorporated into existing lesson plans. Such resources would foster growth and development of these key skills and help connect students to learning in productive ways.

2. Second, as a component of the report card, the CES is communicated to all families four times a year. Thus, it is also important for families to understand the value of the CES and receive supports to cultivate engagement skills at home. Research supports this notion as studies have found that reinforcing school learning and behaviors at home significantly contributes to improved student learning in the classroom (McClelland et al., 2017). To ensure a seamless connection between home and school, supports for the home educators should be aligned with how engagement skills are being discussed and cultivated in the classroom. The goal is to bring teachers and families in concert to build students' knowledge and skills about how to engage with academic tasks and be full participants in collaborative learning with peers.

The CES coupled with teacher supports and parallel family supports would foster partnerships between school and home around the shared goal of supporting students' classroom engagement skills development. Once developed, these home-school supports should be rigorously tested and improved to ensure their efficacy and sustainability in fostering all students' abilities to engage in school. Ultimately, the CES and this system could be extended from pre-kindergarten through third-grade to create a valuable mechanism to measure, monitor, and foster students' development of these skills throughout the critical early grades. Classroom engagement skills previously appeared on the first- grade through third-grade report cards and they were found to be valid and reliable (Barghaus, Fantuzzo, LeBoeuf, Henderson, Li, & McDermott, 2017). A similar measure could be developed for pre-kindergarten as well, to help bridge the critical transition into formal schooling. Having the ability to support students' development of these important skills from pre-kindergarten through third-grade would create an unprecedented system of classroom engagement supports. By doing so, the CES would make visible to teachers, administrators, and families the degree to which our youngest students are connected to their learning activities and the social support systems of the classroom community.

References

Barghaus, K., Fantuzzo, J., LeBoeuf, W., Henderson, C., Li, F., & McDermott, P. (2017) Problems in Classroom Engagement: Validation of an Assessment for District-Wide Use in the Early Primary Grades. *Early Education and Development*, 28(2), 154-166.

Brown, T. A., & Moore, M. T. (2012). Confirmatory factor analysis. In R. H. Hoyle (Ed.), Handbook of structural equation modeling (pp. 361–379). New York, NY: Guilford Press

Chen, F. F. (2007). Sensitivity of goodness of fit indexes to lack of measurement invariance. Structural equation modeling, 14(3), 464-504.

Cheung, G. W., & Rensvold, R. B. (2002). Evaluating goodness-of-fit indexes for testing measurement invariance. Structural equation modeling, 9(2), 233-255.

Cohen, J. (1992). A power primer. *Psychological Bulletin, 112*(1), 155-159.

Data Recognition Corporation. (2015). Technical report for the 2015 Pennsylvania system of school assessment. Retrieved from: http://www.education.pa.gov/Documents/K-12/Assessment%20and%20Accountability/PSSA/Technical%20Reports/2015%20PSSA%20Technical%20Report.pdf

DiStefano, C., Zhu, M., & Mindrila , D. (2009). Understanding and Using Factor Scores: Considerations for the Applied Researcher. *Practical Assessment, Research & Evaluation, 14,* 1-11.

Every Student Succeeds Act of 2015, Pub. L. No. 114-95 § 114 Stat. 1177 (2015-2016).

Fabrigar, L. R., Wegener, D. T., MacCallum, R. C., & Strahan, E. J. (1999). Evaluating the use of exploratory factor analysis in psychological research. *Psychological Methods, 4*, 272–299.

Fantuzzo, J. W., Rouse, H. L., McDermott, P. A., Sekino, Y., Childs, S., & Weiss, A. (2005). Early childhood experiences and kindergarten success: A population-based study of a large urban setting. *School Psychology Review, 34*, 571–588.

Finn, J. D., & Zimmer, K. S. (2012). Student engagement: What is it? Why does it matter? In S. L. Christenson, A. L. Reschly, & C. Wylie (Eds.), Handbook of research on student engagement (pp. 97–132). New York, NY: Springer.

Hu, L. T., & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. Structural equation modeling: a multidisciplinary journal, 6(1), 1-55.

Jones, S. M., Barnes, S. P., Bailey, R., & Doolittle, E. J. (2017). Promoting Social and Emotional Competencies in Elementary School. The Future of Children, 27(1), 49–72.

Kim, E. S., & Yoon, M. (2011). Testing measurement invariance: A comparison of multiple-group categorical CFA and IRT. Structural Equation Modeling, 18(2), 212-228.

Kline, R. B. (2015). Principles and practice of structural equation modeling. Guilford publications.

Liu, Y., Millsap, R. E., West, S. G., Tein, J.-Y., Tanaka, R., & Grimm, K. J. (2017). Testing measurement invariance in longitudinal data with ordered-categorical measures. Psychological Methods, 22(3), 486-506.

Matthews, J. S., Kizzie, K. T., Rowley, S. J., & Cortina, K. (2010). African Americans and boys: Understanding the literacy gap, tracing academic trajectories, and evaluating the role of learning-related skills. Journal of Educational Psychology, 102(3), 757-771.

Millsap, R. E., & Yun-Tein, J. (2004). Assessing factorial invariance in ordered-categorical measures. Multivariate Behavioral Research, 39(3), 479-515.

McClelland, M., Tominey, S., Schmitt, S., & Duncan, R. (2017). SEL Interventions in Early Childhood. The Future of Children, Policy Brief.

McDermott, P. A. (1993). National standardization of uniform multisituational measures of child and adolescent behavior pathology. Psychological Assessment, 5(4), 413–424. doi:10.1037/1040-3590.5.4.413

McKown, C. (2017). Social and Emotional Learning: A Policy Vision for the Future. The Future of Children, Policy Brief.

Muthén, L.K. and Muthén, B.O. (1998-2012). Mplus User's Guide. Seventh Edition. Los Angeles, CA: Muthén & Muthén

Person Inc. (2012). Aimsweb Technical Manual. Retrieved from http://www.aimsweb.com/wp-content/uploads/aimsweb-technical-manual.pdf

Ryu, E., & West, S. G. (2009). Level-specific evaluation of model fit in multilevel structural equation modeling. Structural Equation Modeling, 16(4), 583-601.

Selya, A. S., Rose, J. S., Dierker, L. C., Hedeker, D., & Mermelstein, R. J. (2012). A practical guide to calculating Cohen's f 2, a measure of local effect size, from PROC MIXED. *Frontiers in Psychology*, 3(APR), [Article 111]. DOI: 10.3389/fpsyg.2012.00111

Stapleton, L. M., Yang, J. S., & Hancock, G. R. (2016). Construct meaning in multilevel settings. Journal of Educational and Behavioral Statistics, 41(5), 481-520.

Stott, D. H., McDermott, P. A., Green, L. F., & Francis, J. M. (1988). Learning Behaviors Scale and Study of Children's Learning Behaviors Research Edition manual. San Antonio, TX: The Psychological Corporation.

Waterman, C., McDermott, P. A., Fantuzzo, J. W., & Gadsden, V. L. (2012). The matter of assessor variance in early childhood education—Or whose score is it anyway?. Early Childhood Research Quarterly, 27(1), 46-54.

Appendix A. Methodology

The sections below provide a more detailed description of the methods used in this study. The description is organized by the research question and corresponding method.

Factor Analyses

Research Question 1: Is there scientific support for the CES capturing two distinct dimensions of academic engagement and social engagement?

Factor analytic methods were used to uncover the underlying dimensionality of the CES. To account for student observations nested within teacher raters, two-level models were used for the factor analyses (Stapleton, Yang, & Hancock, 2016). Two-level models are used to partition the variance into that which is explained by the child and by the teacher/classroom and then estimate the factor structure on only the child variance. This was accomplished using a "saturation" method which involves specifying a perfectly fitting factor model at the assessor-level while allowing the child-level to be freely estimated (Stapleton, Yang, & Hancock, 2016; Ryu & West, 2009).

Data were partitioned into two, mutually exclusive subsamples for exploratory factor analysis (EFA) and confirmatory factor analysis (CFA). EFA is used to uncover the number of distinct dimensions that best describe the data based on how the indicators relate to one another. CFA is used to test the fit of hypothesized dimensions to the data. This two-step approach allows for the optimal factor structure to be empirically uncovered using EFA and then cross-validated by testing the fit of the model suggested by EFA to different data using CFA (Fabrigar, Wegener, MacCallum, & Strahan, 1999). For both the EFA and CFA, Mplus version 7 (Muthén & Muthén, 2012) was employed using estimation procedures optimal for categorical data. Model fit was holistically evaluated against multiple research-based criteria outlined below. Finally, unweighted factor scores were generated by averaging the scores across every item associated with a factor from the optimal model derived from EFA and CFA (DiStefano, Zhu, & Mindrila, 2009).

Exploratory factor analysis. A series of two-level EFA models were tested using the EFA subsample to identify the dimensions of the CES based solely on the student variance. For all analyses, weighted least squares (WLSMV) estimation using polychoric correlations was employed and both Geomin and Oblimin factor rotations were tested. The following criteria were used for the model comparison: (a) acceptable model fit indicated by a standardized root mean squared residual (SRMR) less than .08, and *either* a root mean squared error of approximation (RMSEA) less than .08 or a comparative fit index (CFI) greater than .95 following Hu and Bentler's (1999) recommended two-index strategy; (b) the model retains the largest number of items with salient loadings (loadings ≥ 0.40) on only one factor; (c) each factor retains at least four salient items; (d) each factor is internally consistent (r ≥ 0.70); (e) factor are not highly correlated (< 0.85) indicating that distinct constructs exist (Brown & Moore, 2012); and (f) model produces a parsimonious structure aligned with research (Fabrigar, Wegener, MacCallum, & Strahan, 1999). See Appendix B Table 1 for the factor loadings for the

two-factor solution.

Confirmatory factor analysis. Two-level CFA was used to test the factor structures that emerged from EFA, employing the confirmatory CFA subsample. For all analyses, WLSMV estimation using polychoric correlations was employed. The fit of all models was evaluated using Hu and Bentler's (1999) recommendation for a "two-index presentation strategy" in which acceptable model fit is indicated by a SRMR less than .08, and either a RMSEA less than .08 or a CFI/TLI greater than .95 (Hu & Bentler, 1999; Kline, 2015). If a substantial interactor correlation was found (i.e., r > .80-.85; Brown & Moore, 2012), the presence of a general dimension was investigated to determine the unique explanatory contributions of the general dimension as well as any specific factors. This was tested with a confirmatory bifactor model positing one general Engagement factor measured by all items and the specific factors derived from EFA/CFA. Following an evaluation of the model with the best empirical fit, the factor solution chosen was then tested using the data from Cohort 2 to ensure its replicability. Model fit statistics from all confirmatory analyses are provided in Appendix B Table 2.

Measurement Invariance

Research Question 2: Do the dimensions of the CES operate consistently across relevant subgroups of students?

Research Question 3: Are the dimensions of the CES stable across report card marking periods providing an evidence-base for progress monitoring?

The measurement invariance of the CES was assessed to determine if the dimensions of the CES operate consistently across subgroups and time. Multiple group confirmatory factor analysis, using robust estimation to account for the nesting of students within teachers, was used to examine the extent to which the CES dimensions are equivalent across student groups. This analysis assesses the extent to which a factor model's parameters are equivalent across student groups and time which allows for the same model to be used to estimate scores (Kim & Yoon, 2011). Measurement invariance was test for sex, race/ethnicity, English Language Learners, students with disabilities, free/reduced lunch recipients, and from the first to the fourth report card marking period. Race/ethnicity was recoded into four separate dummy variables (with 1 indicating membership in the racial/ethnic group) with white students serving as the comparison group. The Native American/Alaskan Native group was too small (0.13%) for group invariance testing, but was included in the overall sample and concurrent and predictive sub-samples. Measurement invariance across time was assessed between the first and fourth quarters following the procedures outline by Lui, Millsap, West, Tein, Tanaka, and Grimm (2017) for longitudinal data with ordered-categorical measures.

In general for all groupings of interest, a series of models were estimated with cumulative and increasingly more demanding equality constraints on the factor model parameters to test whether the CES operates consistently across groups (Millsap & Yun-Tein, 2004). Each model was estimated in Mplus 7 as a single-level multiple group categorical confirmatory factor analysis model with robust standard errors and chi-squares to account for the clustering of students within teachers. First, a configural model was estimated by constraining the number

of factor and the pattern of item loadings on those factors to be the same across groups. Next, a metric model was estimated in which the factor loadings were constrained to be the same across groups in addition to the constraints imposed by the configural model. Finally, a scalar model was estimated by constraining the item thresholds to be the same across groups in addition to all of the constraints imposed in the metric model (for a detailed discussion of the configural, metric, and scalar model specification see Muthén & Muthén, 1998-2012 and Lui, Millsap, West, Tein, Tanaka & Grimm, 2017).

Traditionally these series of models are compared using chi-square difference tests to determine the extent to which measurement invariance holds. However, with large sample sizes, as is the case in this study, even small deviations in parameter estimates may produce a significant chi-square difference test. Thus, changes in model fit indices, which are less influenced by sample size, were examined to assess measurement invariance (Cheung & Rensvold, 2002; Chen, 2007). Specifically, changes in model fit were used to determine if the more restrictive models (e.g., scalar) fit the data as well as the less restrictive models (e.g., metric), thereby indicating the extent to which a factor model's parameters were equivalent across groups. Models were considered to produce similar fit if the change in RMSEA ≤ .015 and the change in CFI ≤ .01 (Cheung & Rensvold, 2002; Chen, 2007). For model fit statistics for all multiple group models, see Appendix B Table 3.

Student-Level Variance Analysis

Research Question 4: To what degree does a student's score on CES dimensions reflect the student's level of functioning and not extraneous information about the teacher or classroom?

Multilevel modeling was used to identify the proportion of variation in a student's score on the CES that was attributable to influences other than the student (e.g., the teacher of the homogeneous context of the classrooms). Non-student sources of variation in scores can be problematic because it limits the ability of the measure to accurately differentiate children's true abilities. Thus, multilevel models were used to separate the variance in scores on each of the CES factors into two components: (a) variance explained by non-student sources, such as the teacher assessor or classroom (i.e., group-level), and (b) the residual component which is taken to represent the student's ability (i.e., student-level). Using the variance component estimates from this model, the percentage of variance in scores on each factors that is attributable to the student and to the assessor was calculated from the model intraclass correlation [$ICC = \tau 2 / (\tau 2 + \sigma 2)$] multiplied by 100 (Waterman, McDermott, Fantuzzo, & Gadsden, 2012). This approach produces a percentage, ranging from 0 percent to 100 percent, with higher percentages indicating that more non-student variance is captured by the assessment.

Relationships to Other Variables

Research Question 5: Do the identified CES dimensions relate to other important academic and non-academic outcomes measured *concurrently* in kindergarten?

Research Question 6: To what extent do the CES dimensions *predict* other relevant future outcomes?

Evidence based on the relations of the CES to other relevant measures (concurrent and predictive validity) indicates the extent to which CES dimensions are related to existing assessments of similar and dissimilar constructs administered to the same students at approximately the same point in time (concurrent) or in the future (predictive). Below we describe the analytic approach taken to examine concurrent and predictive relations between scores on the CES and other important academic and non-academic outcomes.

Concurrent validity. Seven student outcomes were investigated: number of days absent, number of days suspended, fourth-quarter mathematics grade, and scores on four subtests of the AIMSweb (i.e., Letter Naming Fluency, Letter Sound Fluency, and Phonemic Segmentation Fluency, Nonsense Word Fluency). The AIMSweb outcomes were only available for kindergarten students in Cohort 2 so this cohort was used to test these outcomes. For all seven outcomes, two-level multilevel models partitioned the variance in the kindergarten outcomes into two components: (a) variance explained by non- student sources, such as the teacher assessor or classroom (group-level), or (b) the residual component which is taken to represent the variance attributable to differences in students' engagement behavior (i.e., student-level). For each model, we calculated the percentage of each outcome's student-level variance that could be explained by the CES predictors. Specifically, using the formulation as per Selya, Rose, Dierker, Hedeker, & Mermelstein (2012), we calculated the proportional variance reduction from the null-model compared to the relevant predictive, mixed effects model (i.e., model Psuedo-R2). See Appendix B Table 4 and Table 5 for concurrent validity results from this approach.

Predictive validity. The predictive validity of scores on the CES in kindergarten was examined in two ways. First, we examined the extent to which scores on the CES predict outcomes in third-grade including Attendance, Suspension, and PSSA Mathematics and English Language Arts scores. To do so, a series of two-level multilevel models were estimated for each third-grade outcome. These models separated the variance in the outcomes into two components: (a) variance explained by the clustering of students into schools, (school-level) and (b) the residual component which is taken to represent the variance attributable to the student's behavior (i.e., student-level). For each model, we calculated the percentage of each outcome's student-level variance that could be explained by the CES predictors. Specifically, using the formulation as per Selya, Rose, Dierker, Hedeker, & Mermelstein (2012), we calculated the proportional variance reduction from the null-model compared to the relevant predictive, mixed effects model (i.e., model Psuedo-R2). For predictive validity findings from this approach, see Appendix B Table 6. Second, we examined the extent to which the CES dimensions were predictive of third-grade outcomes above and beyond other kindergarten outcome measures. For this investigation, the following multi-step process was followed for each of the third-grade

outcomes.

1. In the first set of models, a kindergarten outcome was used to predict a corresponding third-grade outcome (i.e., "Model A"). For instance, fourth quarter kindergarten mathematics grades were used to predict scores on the PSSA Mathematics assessment.

2. In a second set of models, a theoretically aligned CES dimension was added to the model predicting the third-grade outcome using the corresponding kindergarten outcome ("Model AB"). For example, the CES Academic Engagement dimension was added to the model predicting scores third-grade scores on the PSSA Mathematics assessment using kindergarten mathematics grades.

3. The Psuedo-R^2 values from Model A and Model AB were then used to calculate Cohen's f^2, an effect size measure of variance explained within a multilevel regression model framework (Selya, Rose, Dierker, Hedeker, & Mermelstein, 2012). Cohen's f^2 indicates how much *additional* variance the CES (Model AB) explained above and beyond the other kindergarten outcome in the model (Model A).

4. The magnitudes of the f^2 values were then evaluated using Cohen's (1992) criteria: small effect indicated by f^2 values between .02 and.15, medium effects between .15 and .35, and large effects by f^2 values above .35). Because f^2 values are scaled to reflect their proportion of variance explained relative to variance explained by the full model they cannot be interpreted directly as a proportion of variance explained (Selya, Rose, Dierker, Hedeker, & Mermelstein, 2012). However, values closer to zero closely match variance explained calculations.

Appendix B. Supplementary Figures and Tables

Figure 1. Sub-Sample Used from Cohort 1 for Different Analytic Purposes.

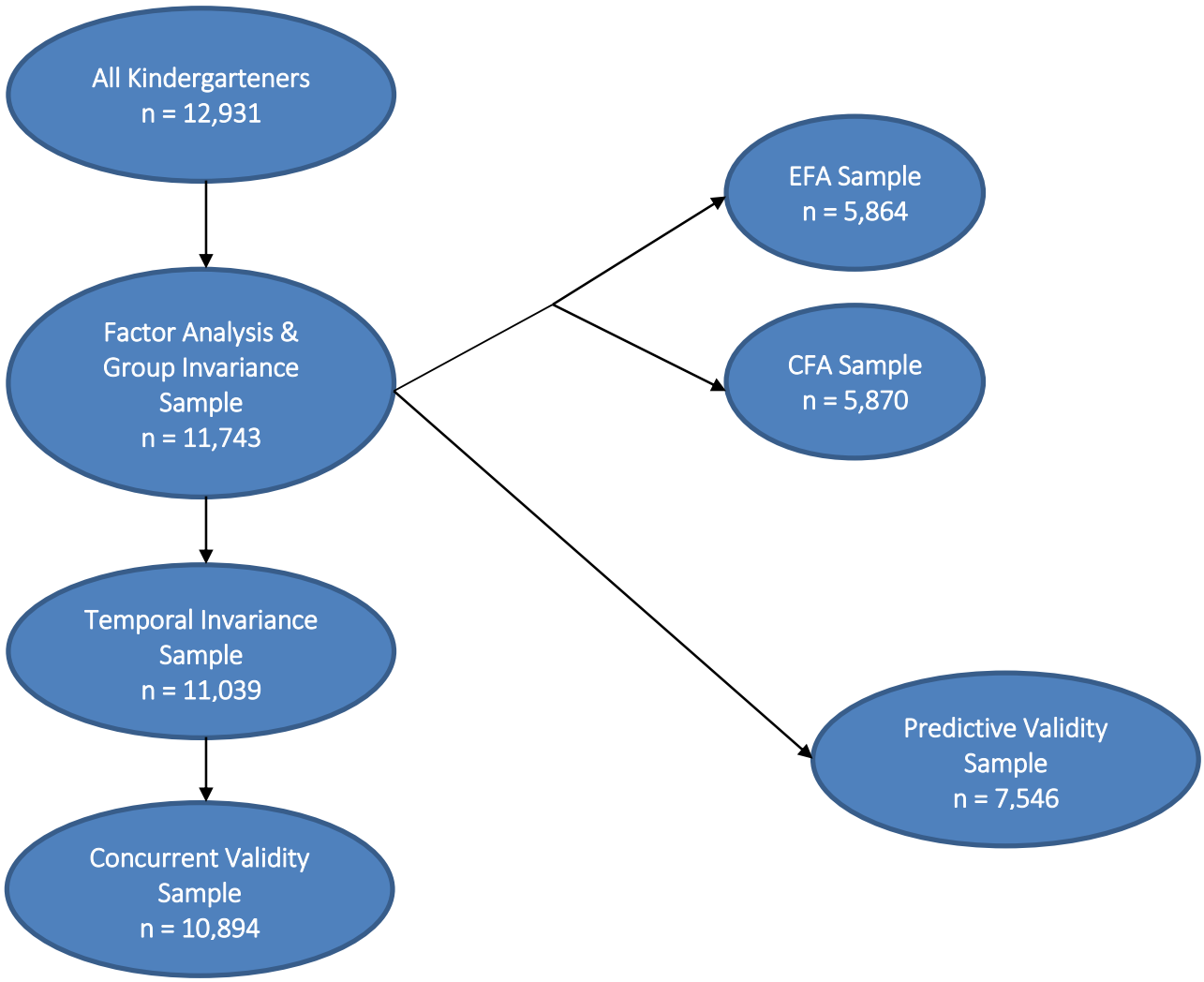


Table 1. Factor Loadings of Classroom Engagement Scale Items on the Engagement Dimensions

Item Prompts	Academic Engagement	Social Engagement
Completes work on time	**1.03**	-0.08
Can work independently	**1.01**	-0.05
Demonstrates consistent effort	**0.78**	0.21
Strives for quality work	**0.77**	0.2
Participates in group activities	**0.62**	0.3
Handles conflict appropriately	-0.08	**1.03**
Respects rights, diversity, feelings & property of others	-0.06	**1.01**
Works and plays cooperatively with others	-0.01	**0.98**
Accepts responsibility for choices and actions	0.01	**0.94**
Makes appropriate movement between activities	0.14	**0.83**
Listens and follows directions	0.14	**0.83**
Respects school environment and materials	0.16	**0.83**

Note. Standardized factor loadings from the exploratory factor analysis are reported.

Table 2. *Model Fit Statistics of the Confirmatory Factor Analysis*

Model Description	χ^2	df	RMSEA	CFI	TLI	SRMR Within
One-Factor Model	17792.09	77	.198	.966	.920	.064
Two-Factor Model	4665.27	53	.122	.990	.974	.026
Bi-factor Model	29265.05	68	.270	.944	.851	.066
Two-Factor Model - Cohort 2	9178.43	53	.125	.987	.967	.024

Note. df = degrees of freedom; RMSEA = root mean square error of approximation; CFI = comparative fit index; TLI = tucker-lewis index; SRMR = standardized root mean square residual.

Table 3. *Model Fit Statistics of the Measurement Invariance Tests*

Model	Description	χ^2	df	RMSEA	CFI	CM	Δ χ^2	ΔCFI	Δ RMSEA
Male	M1: Configural Invariance	5425.2	106	.092	.992				
	M2: Weak Invariance	5454.98	116	.089	.992	M1	51.156	.00	.00
	M3: Strong Invariance	5656.87	126	.086	.992	M2	46.347	.00	.00
Special Education	S1: Configural Invariance	3985.18	106	.079	.993				
	S2: Weak Invariance	3994.32	116	.075	.993	S1	32.453	.00	.00
	S3: Strong Invariance	4122	126	.073	.993	S2	32.104	.00	.00
Dual-Language Learners	D1: Configural Invariance	3643.79	106	.075	.994				
	D2: Weak Invariance	3616.57	116	.072	.994	D1	32.844	.00	.00
	D3: Strong Invariance	3741.49	126	.070	.994	D2	30.931	.00	.00
Free/Reduced Lunch	F1: Configural Invariance	5198.07	106	.090	.992				
	F2: Weak Invariance	5190.8	116	.086	.992	F1	6.089	.00	.00
	F3: Strong Invariance	5374.49	126	.084	.992	F2	9.599	.00	.00
Race (Asian)	Ra1: Configural Invariance	1232.36	106	.091	.995				
	Ra2: Weak Invariance	1226.98	116	.087	.995	Ra1	32.636	.00	.00
	Ra3: Strong Invariance	1278.78	126	.085	.995	Ra2	24.428	.00	.00
Race (Hispanic/Latino)	Rh1: Configural Invariance	1948.03	106	.092	.994				
	Rh2: Weak Invariance	1943.22	116	.088	.994	Rh1	12.974	.00	.00
	Rh3: Strong Invariance	2012.74	126	.085	.994	Rh2	8.273	.00	.00
Race (Multiracial/Other)	Rm1: Configural Invariance	1406.36	106	.098	.995				
	Rm2: Weak Invariance	1404.26	116	.093	.995	Rm1	9.319	.00	-.01
	Rm3: Strong Invariance	1453.97	126	.091	.995	Rm2	4.863	.00	.00
Race (African American)	Raa1: Configural Invariance	3221.05	106	.087	.993				
	Raa2: Weak Invariance	3144.42	116	.082	.993	Raa1	10.394	.00	.00
	Raa3: Strong Invariance	3242.04	126	.080	.993	Raa2	8.614	.00	.00
Temporal Invariance Q1 versus Q4	Te1: Configural Invariance	6598.92	234	.050	.986				
	Te2: Weak Invariance	6742.64	244	.049	.986	Te1	124.47	.00	.00
	Te3: Strong Invariance	6874.8	254	.049	.986	Te2	124.38	.00	.00

Note. χ^2 = target model chi-square. df = degrees of freedom; CFI = comparative fit index; RMSEA = root mean square error of approximation; CM = comparison model; Δ χ^2 = chi square difference test (calculated through DIFTEST Command); Δ change; Weak Invariance = factor loading (metric) invariance; Strong Invariance = factor loading and threshold (scalar) invariance.

Table 4. Variance Explained (R^2) in Outcomes - Cohort 1 Concurrent Validity Sample (n = 10,894)

	Mathematics Grade Fourth- Quarter		Total Absences		Days Suspended	
	b (se)	Psuedo-R^2	b (se)	Psuedo-R^2	b (se)	Psuedo-R^2
Academic Engagement	9.41 (.14)	.31	-4.01 (.23)	.03	-.19 (.01)	.02
Social Engagement	6.23 (.15)	.13	-1.77 (.23)	.01	-.26 (.01)	.04

Note. b = unstandardized regression coefficients; se = standard error; pseudo-R^2 values are calculated as per Selya et al., 2012.

Table 5. Variance Explained (R^2) in Kindergarten AIMSweb Scores - Cohort 2 Concurrent Validity Sample (n = 9,055)

	Letter Naming Fluency		Letter Sound Fluency		Phonemic Segmentation Fluency		Nonsense Word Fluency	
	b (se)	Psuedo-R^2	b (se)	Psuedo-R^2	b (se)	Psuedo-R^2	b (se)	Psuedo-R^2
Academic Engagement	17.30 (.37)	.20	13.79 (.30)	.20	13.62 (.37)	.14	18.64 (.44)	.17
Social Engagement	10.53 (.40)	.07	8.83 (.32)	.08	8.73 (.39)	.05	11.89 (.47)	.07

Note. b = unstandardized regression coefficients; se = standard error; pseudo-R^2 values are calculated as per Selya et al., 2012.

Table 6. Variance Explained (R^2) in Outcomes - Cohort 1 Predictive Validity Sample (n = 7,546)

English Language Arts PSSA			Mathematics PSSA		Total Absences		Days Suspended	
	b (se)	Psuedo-R^2	b (se)	Psuedo-R^2	b (se)	Psuedo-R^2	b (se)	Psuedo-R^2
Academic Engagement	82.38 (2.10)	.17	79.72 (2.19)	.15	-2.84 (.27)	.01	-.38 (.03)	.03
Social Engagement	58.99 (2.13)	.09	55.24 (2.21)	.08	-1.47 (.26)	.00	-.52 (.03)	.05

Note. b = unstandardized regression coefficients; se = standard error; pseudo-R^2 values are calculated as per Selya et al., 2012.