Measuring Social Emotional Learning Through Student Surveys in the CORE Districts: A Pragmatic Approach to Validity and Reliability

Hunter Gehlbach
Heather J. Hough

May 2018

PACE
Policy Analysis for California Education
CORE-PACE RESEARCH PARTNERSHIP

Abstract

As educational practitioners and policymakers expand the range of student outcomes they assess, student perception surveys—particularly those targeting social-emotional learning—have grown in popularity. Despite excitement around the potential for measuring a wider array of important student outcomes, concerns about the validity of the inferences that might be drawn from student self-reports persist. One of the most ambitious attempts to incorporate student perception surveys into a larger assessment framework has occurred through CORE—a consortium of school districts in California. Pulling from CORE's data and their use within these districts, we summarize the evidence for validity and reliability of CORE's student-report surveys on social-emotional learning through a pragmatic approach. After clarifying why validity needs to be viewed as an ongoing process of accumulating evidence (not as an end state), we answer four guiding questions that explain different facets of validity for school leaders:

- How well were the measures designed?
- How well do the measures fit the context?
- With what level of fidelity were the data acquired?
- To what extent are the data being used appropriately?

By detailing the answers to these questions on the student surveys within the CORE districts, we hope to provide guidance around the use of social-emotional learning surveys, both within and outside of the CORE districts. Our ultimate aim is to facilitate decision-making for educational leaders as they weigh decisions regarding the use of student surveys as a component of their assessment programs.

Introduction

As the consensus among policymakers, educators, and the broader public grows around the need for students to develop certain social-emotional skills for a host of academic and life outcomes, research has shown that schools can and do affect the development of these skills (McCormick, Cappella, O'Connor, & McClowry, 2015; Nagaoka, Farrington, Ehrlich, & Heath, 2015). Increasing confidence in schools' roles supporting students' social-emotional development has led some districts and states to include measures of social-emotional learning (SEL) in school accountability systems and continuous improvement plans.[1] This approach is supported by recent federal and state policies, which have encouraged experimentation with different ways to incorporate SEL into school performance measurement systems. The 2015 Every Student Succeeds Act (ESSA) requires that states measure at least one non-academic indicator of "school quality or student success" (e.g., student engagement, post-secondary readiness, or school climate and safety), opening up the opportunity for SEL measurement (LaRocca & Krachman, 2017). Similarly, under California's Local Control Funding Formula and the supporting Local Control Accountability Plan, districts are expected to develop and report indicators representing a wide range of educational goals, including school culture and climate. As schools and districts focus on students' SEL, and the school conditions that support it, there is growing demand for measures that track students' progress over time. While there is a variety of ways to measure SEL, including performance assessments and embedded tasks,[2] student self-reported surveys are of growing interest, as they are relatively inexpensive to collect. However, even as demand for these measures grows, so do concerns that these surveys may not provide reliably actionable information about students or schools. To help navigate this tension, we outline the emerging evidence of validity and reliability for one SEL survey, developed by the CORE districts, and highlight the primary issues and questions that should be addressed for educational leaders and policymakers considering the use of survey measures such as these in school performance measurement systems.[3]

The CORE districts (Fresno, Garden Grove, Long Beach, Los Angeles, Oakland, Sacramento, San Francisco, and Santa Ana) together serve nearly one million students. Their systematic measurement of school and student performance at scale began when these districts received a "waiver" from the U.S. Department of Education. This waiver freed six of the CORE districts[4] from certain No Child Left Behind (NCLB) obligations and enabled them to develop innovative approaches to school accountability. To support their vision of educating

[1] While no state has chosen to measure SEL as part of their ESSA plan at this time (Blad, 2017), in 2017, all 50 states had SEL standards at the preschool level, and eight states had SEL standards for K–12 (Dusenbury, Dermody, & Weissberg, 2018). Additionally, many more states are working to build capacity in developing approaches to measuring and improving SEL. For example, 25 states are currently working with the Collaborative for Academic and Social and Emotional Learning (CASEL) through the Collaborating States Initiative (R. Weissberg, personal communication, March 23, 2018).

[2] See, for example, https://measuringsel.casel.org/.

[3] Other researchers are working to document validity and reliability evidence across all available SEL measurement approaches. See, for example, Schweig, Hamilton, Stecher, and Baker (2017).

[4] Garden Grove and Sacramento Unified school districts were not part of the waiver but are part of the CORE network.

the "whole child," under the terms of the waiver, the participating districts developed a multiple-measures data system that better reflected their vision for school and student performance. This system—along with multiple measures of student academic performance and growth and non-academic outcomes such as attendance and school climate—included measures of social-emotional learning based on student self-report surveys. The CORE districts' SEL survey comprises a battery of items designed to measure four SEL constructs: self-management (9 items), social awareness (8 items), growth mindset (4 items), and self-efficacy (4 items). Students in Grades 3 through 12[5] rate themselves on the same 25 questions using a 5-point Likert scale. The four SEL constructs are defined as follows:

- **Self-management**, also referred to as self-control or self-regulation, is 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).
- **Growth mindset** is the belief that one's abilities can grow with effort. Students with a growth mindset believe that they can develop their skills through effort, practice, and perseverance. These students embrace challenges, see mistakes as opportunities to learn, and persist in the face of setbacks (Dweck, 2006).
- **Self-efficacy** is the belief in one's ability to succeed in achieving an outcome or reaching a goal. Self-efficacy reflects confidence in the ability to exert control over one's own motivation, behavior, and environment and allows students to become effective advocates for themselves (Bandura, 1997).
- **Social awareness** is 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).

These surveys were designed to aggregate students' individual reports to create school level scores (which could be disaggregated by subgroup). Under the waiver, the SEL surveys were assigned weights and used in conjunction with the other metrics to create a single school composite that was then used to identify schools for improvement.[6] With the passage of ESSA, the CORE districts are no longer beholden to the terms of the waiver, thus, they no longer worry about creating composite school scores or attach consequences to the SEL student surveys. However, the districts still administer SEL surveys to students in Grades 3–12 and use them in a multiple-measures framework to understand school and student performance in the context of their work together as a Networked Improvement Community.[7] CORE remains the

[5] All districts collect the survey for students in Grades 4–12; only a subset collects the surveys from third-grade students.

[6] The other metrics in CORE's measurement system are: academic achievement and growth; suspension/expulsion rates; chronic absenteeism; high school readiness; graduation rate; English Learner redesignation; and school culture-climate reports from students, staff, and parents. For more detail on the use of CORE's measurement system under the waiver, see Marsh, Bush-Mecenas, Hough, Park, Allbright, Hall, & Glover (2016).

[7] For more on CORE's current work as a Networked Improvement Community, see Nayfack, Park, Hough, & Willis (2017).

only education system to measure SEL at scale, and as such continues to attract widespread national interest in the field of education and in the popular press (e.g., Guzman-Lopez, 2017; Zernike, 2016).

Because SEL measurement at scale is so new nationally, studying the properties of these measures and their use is a central focus of the research coordinated by the CORE-PACE Research Partnership. In this brief, we summarize this existing body of work within CORE to provide educational leaders and policymakers with a way to think through the use of student-report SEL surveys. We begin by providing a framework for how to approach "validity," including reliability as a component of validity. We then draw from our work in the CORE districts to illustrate what we know about the validity of CORE's SEL measures and identify key areas that require further study. We conclude with a summary of the key considerations for educational leaders and policymakers considering SEL measurement and use.

A Pragmatic Approach to Validity

Although numerous schools and districts survey their students about aspects of their social-emotional learning, the practice raises concerns from critics and champions alike (e.g., Duckworth & Yeager, 2015). Fundamentally, most of these worries revolve around reasons why a particular measurement might not be credible in a particular context. These "threats to validity" come in a host of forms with numerous technical appellations. We propose four basic questions to facilitate the evaluation of these threats and, conversely, weigh the evidence in favor of validity.

To understand how these questions speak to the validity of a given measure, we begin by conceptualizing validity as a process (Gehlbach, 2015). One commonly hears the challenge, "But has this measure been validated?" from policymakers, district data administrators, and academics alike. Yet, a "validated scale" is a mythical entity. Instead, validity is an ongoing process that begins with the purposeful development of a measure (Gehlbach & Brinkworth, 2011), entails accumulating evidence of that measure's characteristics over time (Messick, 1995), and relies upon logical arguments that draw from that body of evidence to make the case that the measure assesses what it purports to in a particular context for a particular population and for a particular use (Kane, 1992, 2006). In many ways, this process of validity resembles the work of a trial attorney—lawyers must purposefully develop a logical argument using available evidence; that evidence might grow or become reinterpreted in light of new evidence over the duration of a trial. Thus, this brief will not describe whether the social-emotional learning scales used by the CORE districts are valid, but rather how compelling the scales' evidence of validity is currently. Because more data are collected annually and research on new measures is an ongoing process, the amount of evidence for validity changes over time.

We describe our approach to validity as ***pragmatic*** because we recognize that school leaders frequently face constraints different from those of academic researchers. Schools cannot devote infinite amounts of time to surveying students; students' tolerance for completing such measures is notably finite; what scales are included within survey measures can be politically contentious in local communities; how schools report comparisons from the survey data of different student groups can be highly contentious; and so forth. As a result, schools are forced to make trade-offs around the validity of any survey measures they might wish to use. We hope that this brief provides some preliminary guidance for school leaders thinking through some of the more important trade-offs.

With this definition of validity as a process, we can unpack our four focal questions that address different types of validity (summarized in Table 1). Because our ultimate goal is to focus on the social-emotional learning scales that CORE has used to measure the underlying constructs of self-efficacy, growth mindset, social awareness, and self-management, we constrain our discussion to the validity of survey measures in general and these surveys in particular (though many of the same principles apply to test scores and other measures).[8]

How Well Were the Measures Designed?

One large set of validity concerns pertains to questions around how well the measures were designed (Gehlbach & Brinkworth, 2011; Messick, 1995). First, measures must be designed for ***reliability***. If a survey scale suggests that students' growth mindsets vary wildly over the course of a week (a time span in which growth mindsets are presumably very stable), the scale would lack reliability and, consequently, making a case for validity would be challenging.

Another key characteristic is ***content validity***: Did the survey designers include the right questions to measure the focal topic (i.e., the underlying construct)? For example, to measure math self-efficacy, one might construct items about performing well on tests, understanding the teachers' lectures, and solving hard problems. However, one should not include questions about enjoyment of math class because enjoyment is a different construct (and is therefore the wrong content to include in a self-efficacy scale).

In addition, items should have ***face validity***, i.e., at face value, items should clearly signal the construct they are purporting to measure. Asking about someone's political orientation might predict their environmental attitudes, but, at face value, asking about political orientation does not seem like a valid approach to assessing environmental beliefs.

Researchers often say that a survey has ***structural validity*** if the items within the survey actually align with the construct(s) that they were supposed to assess in the pattern or structure that was expected. In other words, if a set of items that was designed to measure a single construct actually contained multiple constructs, it would demonstrate evidence of a lack

[8] Note that the CORE districts also administer school culture-climate surveys to all students, parents, and school staff. While validity of these surveys is important to establish as well, here we focus on CORE's SEL surveys, because CORE's culture-climate surveys draw heavily from the California school climate surveys, which are further described elsewhere: California Healthy Kids Survey (http://chks.wested.org/), California School Climate Survey (http://cscs.wested.org/), California School Staff Survey (http://csss.wested.org/), and California School Parent Survey (http://csps.wested.org/).

of structural validity. Likewise, if a set of items was designed to assess two distinct constructs but only showed evidence of one, it would be a bad sign for the structural validity of the measure. Note, however, that a survey construct can contain items at different levels of abstraction to try to create multiple, specific scales of a single overarching scale. For example, CORE describes *social awareness* as “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” (West, Buckley, Krachman, & Bookman, 2017). Thus, while the designers of CORE’s survey decided to create just one measure of “social awareness,” they could have just as easily designed the survey to try to obtain distinct measures of social perspective taking, ethical behaviors, and resource recognition, which would have resulted in a very different structuring of this larger construct. For most researchers, what matters is that the indicators within a scale all signal that a single construct is being assessed, and there are statistical techniques can that assess this.

Additionally, items on a scale should be ***representative*** of the domain they are designed to measure. In other words, scales should be designed to include a broad cross-section of indicators of the construct. Returning to the earlier example of self-efficacy, in addition to items on test performance, teachers’ lectures, and hard problems, one might argue that an item about confidence in doing math homework might also be important to adequately represent the entire construct.

Finally, survey items need to be designed in ways that adhere to ***best practices*** that will mitigate measurement error (Dillman, Smyth, & Christian, 2014; Gehlbach, 2015). Common practices such as inclusion of double-barreled items (where survey designers inadvertently embed two questions in a single item, such as, “How important is it for you to be rich and happy?”) are known to infuse measurement error into scales thereby reducing the validity of the measure.

Historically, most of the attention in designing survey scales has focused on post hoc statistical analysis of items that assesses aspects of validity such as reliability and structural validity (DeVellis, 2003). Recently, more attention has been given to thinking about validity from the outset of the design process (Gehlbach & Brinkworth, 2011). For instance, techniques such as an expert review (Rubio, Berg-Weger, Tebb, Lee, & Rauch, 2003) can help bolster the validity of the content of a survey scale. Likewise, increasing awareness of an entire science of survey design—one that can offer guidelines to minimize measurement error (Dillman et al., 2014; Fowler, 2009)—can help to improve the validity of measures.

How Well Do the Measures Fit the Contexts?

This second aspect of validity addresses the fit between the measures and the contexts in which they are supposed to function. Here we ask, does this survey work well in certain schools or grade levels or when administered in certain ways, but not in others? One potential concern around the interaction between the measure and the context is whether ***floor or ceiling effects*** occur. For example, a scale with problematic ceiling effects might fail to differentiate students in a school where almost everyone has strong self-management competencies, even if the scale successfully differentiates students in a school where most students have low self-management skills. This would be a problem because the items in the scale are not sensitive enough to distinguish students with strong self-management skills from those with *very strong* self-management skills.

A related threat to validity is ***reference bias*** (West et al., 2016), which occurs when students in one context may respond to surveys very differently from students in a different context based upon the peer norms that they observe. In other words, the same student might report having very high or very low self-management skills depending on their perceptions of how organized their peers are. Thus, measures of relatively overt behaviors (like self-management) would typically be more likely to be vulnerable to this bias than measures of internal psychological processes (such as growth mindset), since students do not directly observe internal process and thus cannot compare others' to their own.

In education, school performance measures are expected to allow comparisons between subgroups of students. Consequently, in order for a survey to demonstrate validity, different subgroups of students need to interpret and react to the same items in the same way. However, if students in different subgroups answer questions differently, scales would demonstrate ***measurement invariance,*** which would be evidence against a scale's validity. Threats to validity here can occur if the survey has different meanings for different respondents (e.g., some students might perceive an item that asks about "jumping into the class conversation without being called on" as a virtue, while others view it negatively) or even if items are reinterpreted over time by the same student (e.g., a student who recalibrates what it means to be "good at math" after beginning algebra, which suddenly seems much more challenging). Additionally, items or scales with cultural bias would likely cause problems with measurement invariance, if students in different racial groups interpret questions differently based on cultural norms or experiences. The extent to which this potential problem exists for a particular scale can be tested qualitatively through cognitive interviewing (Willis, 2005) or quantitatively through measurement approaches such as those under item response theory (IRT) or structural equation modeling.

Perhaps because there are an infinite number of contexts that any given measure might be applied to, knowledge surrounding the "fit" characteristic of validity seems not as well developed. Some issues, such as whether to be concerned about floor or ceiling effects, can easily be informed by results from pilot studies. However, other issues regarding how well the measures fit the context are less easily resolved. For example, while issues of reference bias and problems of measurement invariance have been documented (e.g., West et al., 2016), how frequently they occur and how problematic they are for pragmatic decision-making is less clear. In many cases, the best ways of trying to optimize validity by ensuring that the measures fit the context are to rely on researcher judgment. Knowledge about how evidence of validity for a particular context might transfer to a different context will presumably emerge as researchers begin to address this question more directly.

With What Level of Fidelity Were the Data Acquired?

Although this issue is not typically addressed in academic conversations around validity, ostensibly good measures might produce bad data through a host of problematic administration procedures and/or respondent practices. Too little is known about which ***survey administration practices*** optimize data quality. However, it seems likely that an apathetic teacher who tells students to take a survey because the administration is forcing him to do so while rolling his eyes will receive a different caliber of data than the counselor who impresses upon students how valuable the data are for the school's decision-making processes and how much she and other administrators value students' voices.

One concern when students are taking surveys is the extent to which they may employ a host of different ***satisficing*** strategies (Krosnick, 1991)—ways respondents avoid putting forth effort while completing surveys. These strategies include: rushing; skipping items; stopping the survey without answering all of the answers; anchoring-and-adjusting, in which a respondent answers subsequent questions based on their responses to the first (Gehlbach & Barge, 2012); straight-line responding, where students give the same answer (e.g., the second response option) for every question (Barge & Gehlbach, 2012); and mischievous responding (Robinson-Cimpian, 2014). To the extent that respondents want to present themselves in a favorable light, ***social desirability bias*** may also cause respondents to give the answers they think would impress others rather than truthful responses. Depending on the length of the survey, fatigue could also influence data quality. Thus, for school administrators, the fidelity of data collection is another important component of validity.

How seriously respondents engage with a survey has outsized effects on the validity of the resultant data. Like many of the issues in the previous section, much of what we know about improving data quality through survey administration procedures more closely resembles common sense than empirically derived guidelines. However, a number of new practices (e.g., recording how quickly respondents complete web surveys) can help to detect satisficers of various types (Barge & Gehlbach, 2012). By contrast, socially desirable responding is likely to be harder to detect but might be easier to avoid. For example, survey designers might revise questions to avoid any potential embarrassment on the part of respondents (Tourangeau & Yan, 2007).

To What Extent Is the Data Being Used Appropriately?

This final set of validity concerns revolves around the use of the scores that are derived from the survey measures. One of the first considerations in this domain of validity is whether data ***analysis and measure creation*** are congruent with the expected use(s) of the data. Once survey data have been collected, many analytic decisions need to be made: Response options can be treated as ordinal or continuous, Item Response Theory (IRT) or factor analytic solutions might be employed to create more sophisticated scales, weighting of responses might be used in an attempt to get closer to a representative sample, or data may be aggregated in different ways. These analytic decisions can have major implications for how the survey measures are interpreted and used.

Among the most frequently discussed forms of validity in this domain are: ***convergent validity*** (do high scores on one measure correspond with high scores on similar measures?), ***discriminant validity*** (are scores on a measure uncorrelated with scores on unrelated measures?), and ***predictive validity*** (do scores on a measure allow us to anticipate future outcomes?).

In thinking about adopting or using a new measure, school officials also have to wrestle with how ***generalizable*** they think a particular set of findings might be. The broad umbrella of questions that generalizability addresses is whether results from one context can be generalized to a different setting or domain. For example, will results from this year's fifth graders' self-management skills give us a sense of next year's fifth graders? Do self-efficacy scores in science generalize to math? Can an instrument developed in one school district be used in another?

Another major concern for decision-makers is understanding the consequences (intended and unintended) of how scores are being used, also referred to as ***consequential validity***. Survey scores that are used to evaluate teachers or determine their salary versus providing formative assessments versus screening individual students to determine who is at risk will inevitably have distinct consequences. These consequences, in turn, may affect the validity of survey scores themselves (e.g., students trying to game a high-stakes survey to get their teacher in trouble). Relatedly, recent research suggests that surveys can function not only as data collection devices but also as ***interventions*** in and of themselves (Gehlbach, Robinson, Finefter-Rosenbluh, Benshoof, & Schneider, 2018): The mere act of measuring can direct people's attention and change behavior.

Like the first validity domain, there is a robust empirical base surrounding many of these aspects of validity. Whole journals (and corresponding scholarly debates) have been dedicated to structural equation modeling and IRT—both common, yet different, approaches to examining survey data. A number of conventions guide approaches to establishing convergent, discriminant, and predictive validity (Kenny, 1995). Issues where one typically must be guided by intuition include: how to judge how well one's findings might generalize, how to anticipate the unintended consequences of a particular use of survey data, or how to realize when a survey is turning into an intervention that is affecting data quality.

Table 1. Four Questions and Their Components to Guide a Pragmatic Approach to Validity

How well were the measures designed?	
Reliability	Does a scale produce consistent scores over time when no changes have actually occurred?
Content validity	Does each survey scale cover appropriate indicators of that topic or construct? (And do they omit indicators of related but distinct constructs?)
Face validity	Do the items within each scale appear to measure the construct they are supposed to measure? In other words, if taken at face value, do the items seem like appropriate indicators for that scale?
Structural validity	Does a scale that was designed to represent a single construct represent only that one construct or do its psychometric properties suggest it is measuring more than one construct?
Representativeness	Do the items within a scale provide a representative cross-section of that construct?
Best practices	To what extent are survey items written in ways that adhere to best practices and thereby minimize measurement error to the extent possible?
How well do the measures fit the context?	
Floor/ceiling effects	How well do the items on a scale spread respondents out across a range of responses for a particular population?
Reference bias	To what extent do respondents answer a survey scale differently based on the local peer norms?
Measurement invariance	Does one subpopulation of respondents interpret the items of a survey scale as meaning the same thing as does a different subpopulation?
With what level of fidelity are the data acquired?	
Survey administration practices	How well does the survey administration motivate students to answer each item to the best of their abilities?
Satisficing	To what extent do students engage in strategies to avoid putting effort into completing the survey with fidelity?
Social desirability bias	To what extent are students trying to present themselves in a favorable light as they answer (rather than striving to provide truthful responses)?
To what extent are the data being used appropriately?	
Analysis and measure creation	How appropriate are the decisions made during measure creation, data analysis, and reporting given the intended use of the survey data?
Convergent and discriminant validity	How well does a measure correlate with other, related measures that it should theoretically correlate with? How well does a measure demonstrate a lack of statistical relationships with measures that it is theoretically distinct from?
Predictive validity	How accurately does a measure predict future outcomes that we might care about?
Generalizability	Given the nature of the construct and the nature of the population, to what extent are the results likely to generalize to other related constructs and/or other populations of respondents?
Consequential validity	What are the intended and unintended consequences of a particular use of survey scores?
Surveys as interventions	The mere act of measuring can change people's attention, priorities, or behavior. To what extent does a survey (intentionally or unintentionally) serve as an intervention? To what extent does such an intervention compromise the quality of the data that are collected?

Evidence for the Validity of CORE's SEL Measures

With these broad categories of validity in mind, we now apply them to examining the robustness of CORE's SEL measures within each domain. In this section, we bring to bear evidence of validity from all of the research that has been conducted to date using CORE's SEL data. Most analyses use CORE's first two years of administration of the SEL survey at scale (2014–15 and 2015–16) of students in Grades 3–12 in five districts,[9] in which 390,000 (about 70 percent) of students in the districts took the survey in each year. Some analyses draw from a pilot administration in 2013–14, in which some districts also surveyed teachers about individual students' social-emotional learning.[10] Students attending schools in the CORE districts in these grades are diverse: 69 percent are Latinx, 10 percent are African American, 7 percent are Asian, 73 percent are economically disadvantaged, and 36 percent are classified as English language learners (West, Pier, Fricke, Hough, Loeb, Meyer, & Rice, 2018).

How Well Were CORE's SEL Measures Designed?

Within CORE, the survey design process began with substantial thought around which SEL constructs to measure and why they were being measured (Krachman, Arnold, & LaRocca, 2016; Allbright, Marsh, & Hough, 2017)—arguably the most important part of any survey design process. The survey designers focused on selecting SEL constructs that were "meaningful, measurable, and malleable," and balanced internal stakeholder priorities with external guidance by convening district representatives and SEL experts. Ultimately, these individuals prioritized constructs by voting. It is worth noting that although they involved numerous people in the process and solicited a wide array of opinions on what to measure, the survey designers avoided the common pitfall of allowing different individuals to suggest items they happen to like and instead focused on identifying coherent scales to measure key underlying constructs (Gehlbach & Artino, 2018). This group also put extensive thought into the larger context that these SEL assessments would fit into, which suggests attention to consequential validity from the outset of their design process. Through a series of pilot tests, CORE assessed the face validity of their measures with educators and engaged in ongoing conversations with content experts for each of the measures. They complemented these conversations with discussions of best practices in item wording with other experts. These conversations allowed the survey design team to address concerns around ***content validity***, ***face validity***, ***representativeness of items,*** and ***best practices*** in wording items from early on in their design process.

The resulting data from their initial administrations then allowed CORE to assess the structural validity and reliability of each scale through a variety of statistical techniques. These analyses provided the opportunity to make informed decisions about how to adapt the scales,

[9] One CORE waiver district collected student data anonymously, so it is not included in student-level analyses or analyses that require tracking students over time.

[10] For more detail on CORE's development of the SEL surveys, pilot study, and survey rollout, see Krachman, Arnold, & LaRocca (2016).

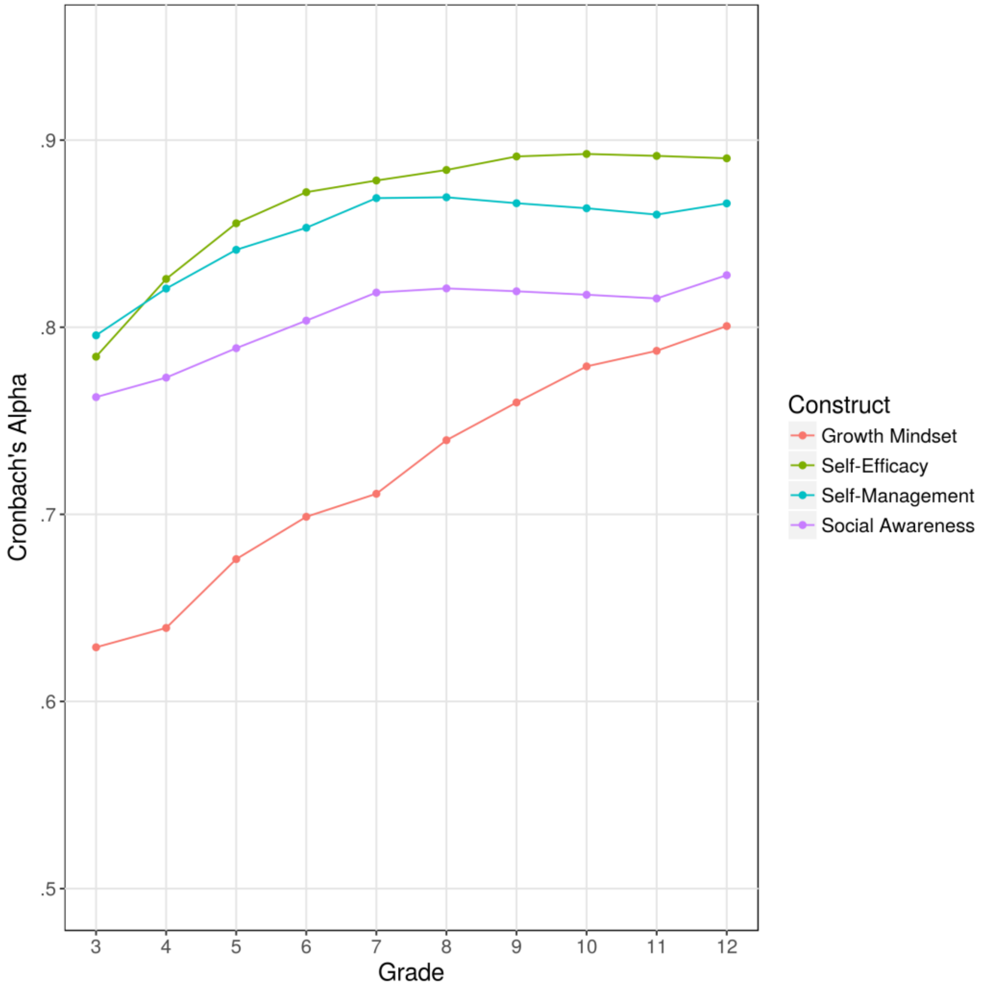
in the pilot phase.[11] Through ongoing research, the districts continued accumulating evidence of validity and identifying areas of improvement.[12] For instance, a recent analysis demonstrates that the ***structural validity*** of the four scales is sound as assessed through confirmatory factor analysis and IRT methods. In other words, the items on the survey appear to be measuring distinct, separate constructs (Meyer, Wang, & Rice, 2018).

Researchers have also found that the ***reliability*** of the scales is strong for almost all scales, although the growth mindset scale appears to be less reliable for younger students—particularly those in elementary school (Meyer et al., 2018). Figure 1 shows the reliability as measured by Cronbach's alpha (a measure of how consistently students answer the items within a construct), for each SEL construct and at each grade level. The reliability of the self-management, self-efficacy, and social awareness scales is all comfortably higher than the frequently recommended .70 level (DeVellis, 2003), ranging between .77 and .89, although the reliability coefficients are generally higher at higher grades. However, the reliability coefficients of the growth mindset scale, especially below Grade 7, are lower than .70, indicating that the data from these younger students contain less signal and more noise. All four growth mindset items are negatively phrased, and there is some speculation that the negatives in many of these items (e.g., "If I am not naturally smart in a subject, I will never do well in it.") may pose cognitive challenges for these younger students (see also Benson & Hocevar, 1985; Gehlbach & Artino, 2018). However, CORE researchers have found success through IRT mixture models to mitigate the challenges associated with negative wording (Bolt, Wang, Meyer, & Rice, 2018). In addition, via subscore augmentation techniques, which allow for incorporating collateral information from the entire SEL survey, the researchers have shown the potential of increased reliability of SEL scores (Meyer et al., 2018).

[11] For more on the pilot testing and development of CORE's SEL surveys, see West, Buckley, Krachman, and Bookman (2017).

[12] All current and future work can be found at http://www.edpolicyinca.org/projects/core-pace-research-partnership.

Figure 1. Cronbach's Alpha Coefficients of the SEL Constructs at Each Grade Level



Source: Meyer, Wang, & Rice (2018).

In sum, CORE has established good evidence of these initial design aspects of validity. Nevertheless, there is still much to learn. For instance, internal reliability on these measures is high (as assessed by Cronbach's alpha), particularly when contrasted with standardized achievement tests, which typically rely on many more items to ensure reliability. However, other kinds of reliability, such as test–retest, have not yet been fully explored. Recent work shows that SEL measures are much less correlated across grades than academic test scores, even when corrected for measurement error (West, Pier, et al., 2018). Is this a result of instability in responses to the measures across time? How different are students' responses if questioned at multiple points of the school year? Do they answer differently before or after a big test? These types of questions require additional attention.

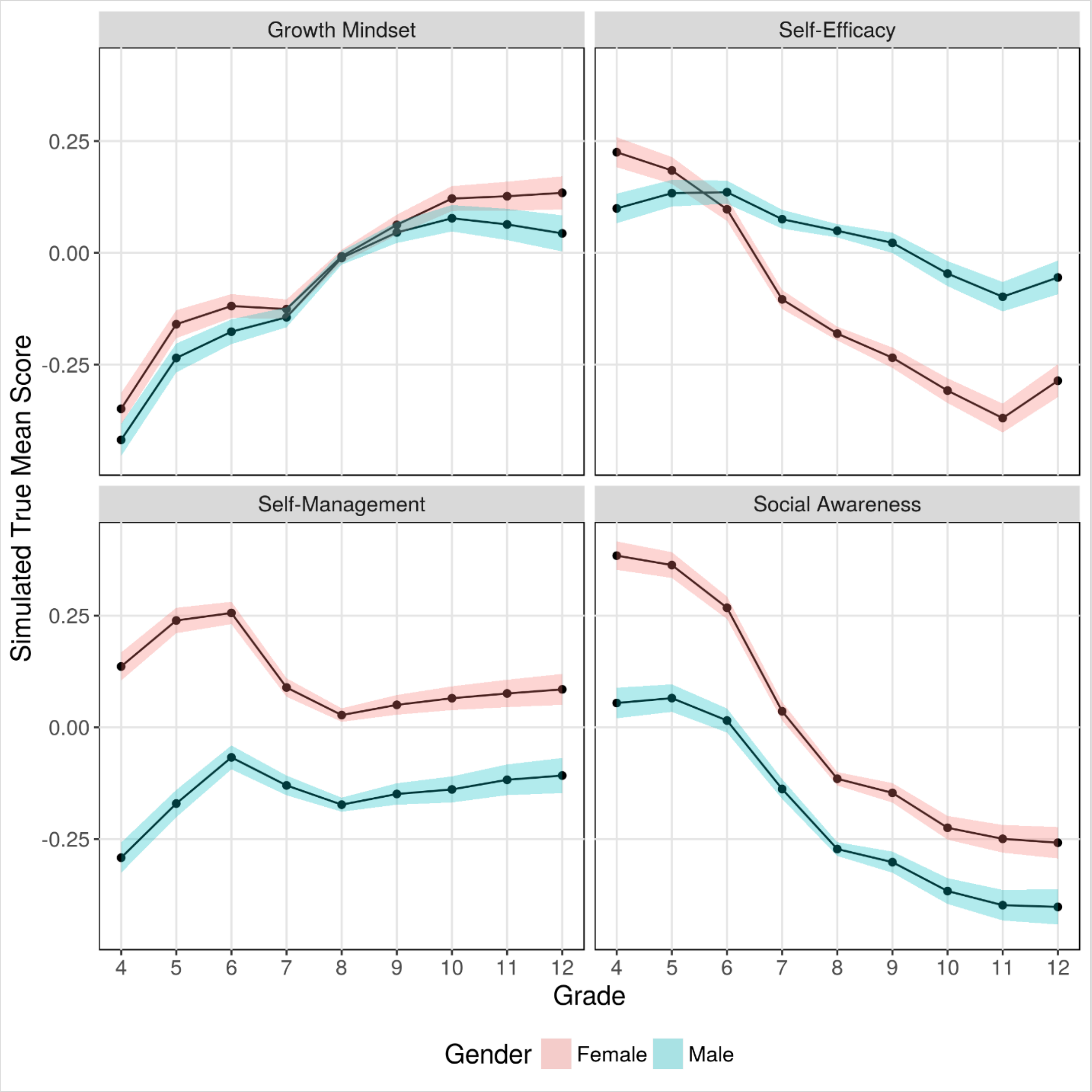
Furthermore, most of what is known about best practices in designing and wording survey items comes from experiments on adults (e.g., Krosnick & Presser, 2010). Because the cognitive sophistication of elementary and secondary students will typically be lower, it is unclear how applicable these best practices are for younger respondents. Additionally, more systematic inclusion of certain steps in the survey design process such as expert reviews or cognitive pretesting could provide additional, compelling evidence for validity (Gehlbach & Brinkworth, 2011).

How Well Do the Measures Fit the Diverse Contexts That Make Up CORE's Schools?

In the years since the first large scale administration of CORE's SEL surveys in 2014–15, the measurement team has invested substantial work in understanding how the measures are functioning for the students and schools that are served by the CORE districts. For example, because each item on the surveys is presented on a scale of 1–5, **floor/ceiling effects** are a potential threat to validity. As Meyer et al. (2018) have shown, the SEL items have reasonable spread, which helps to distinguish between students with high and low social-emotional skills. In addition, IRT scale scores can further mitigate any skew in raw score distributions. (Discussion of measurement approaches using the CORE data is expanded in a later section).

With survey administration in such a diverse context as the CORE districts, an important consideration in the use of the survey measures is in ensuring that all students are interpreting and responding to the questions in the same way, or that there is no evidence of ***measurement invariance***. This is particularly important because we see large differences in student responses across grades and student subgroups (West, Pier, et al., 2018). For example, Figure 2 highlights the differences we see across grades and between students of different genders. We see that students respond quite differently to the surveys as they progress through the grades, and that boys and girls have very different reports on nearly every construct (West, Pier, et al., 2018). We want to be sure that these reflect real differences in the underlying constructs, not differences in interpretations of the questions.

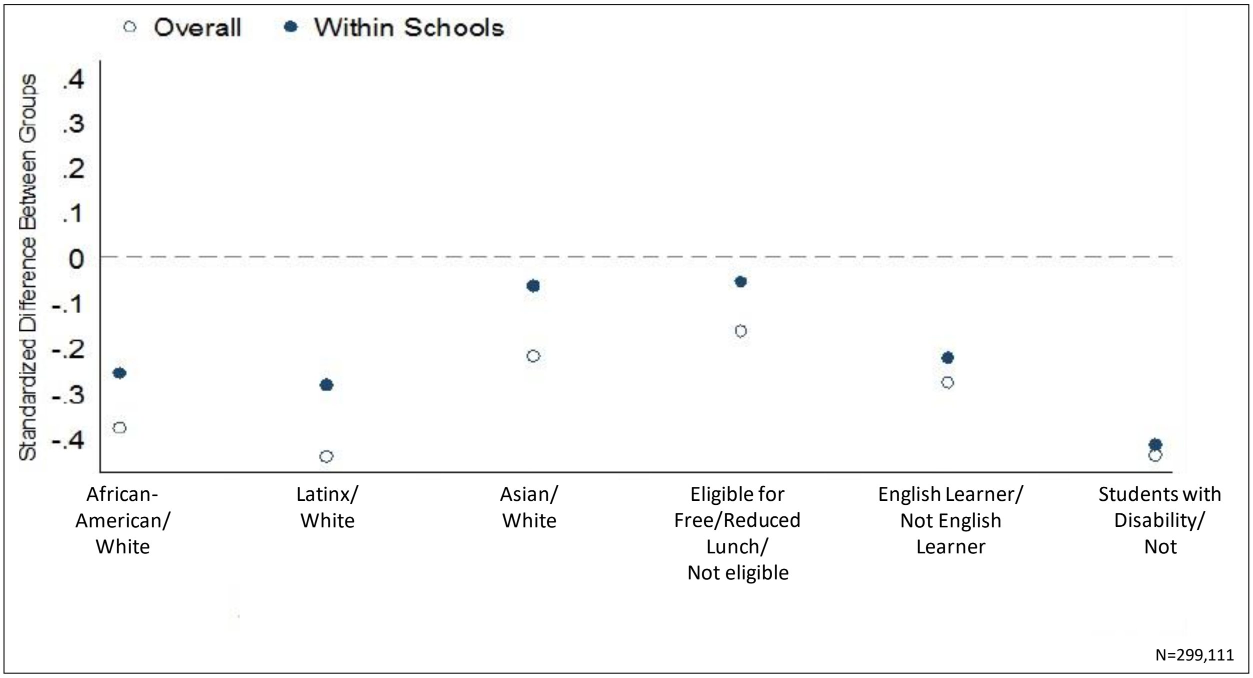
Figure 2. Trends in Mean SEL Construct by Gender, 2015–16



Source: West, Pier, Fricke, Hough, Loeb, Meyer, & Rice (2018).

Similarly, we see substantial and consistent gaps by student racial/ethnic group. As shown in Figure 3, students in special education, African American students, and Latinx students report the lowest levels of SEL, and differences between these groups persist even within schools. In Figure 3, a score of zero indicates no difference between groups, with those groups furthest from the line demonstrating the largest gaps. For example, Latinx students report an SEL score that is 0.36 standard deviations lower than White students even after controlling for other demographic characteristics. Comparing students within the same school, the gaps are smaller, but still substantial (0.24 standard deviations lower than White peers in the same school) (Hough, Kalogrides, & Loeb, 2016).

Figure 3. SEL Gaps by Student Demographics, Overall vs. Within Schools



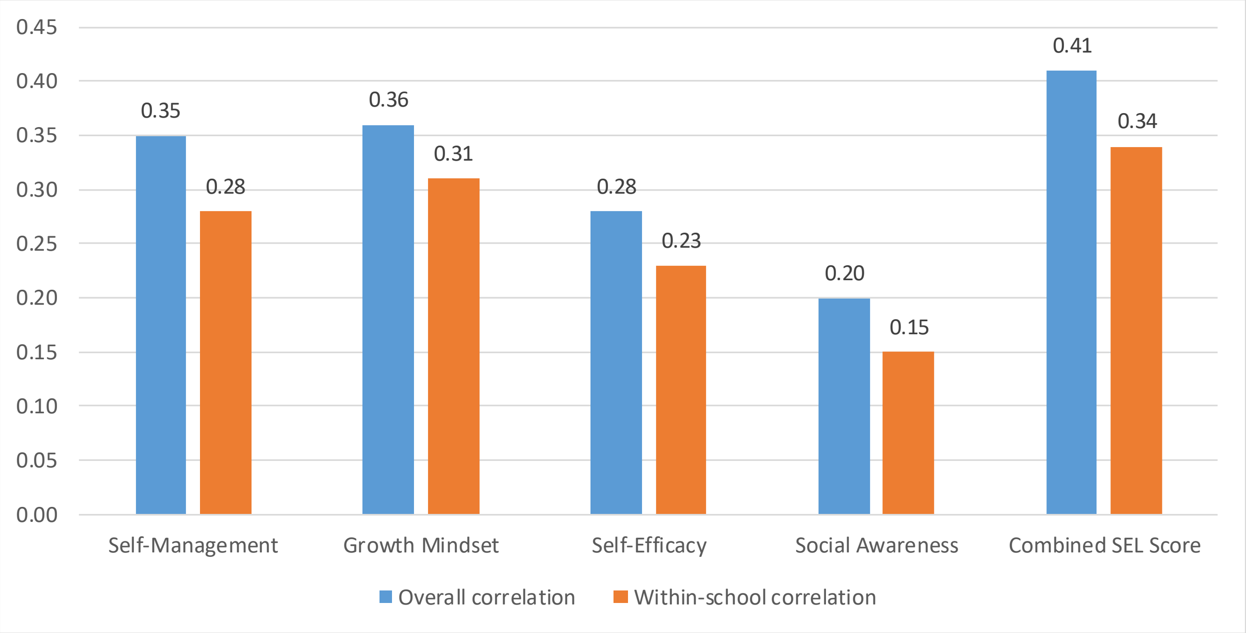
Source: Hough, Kalogrides, & Loeb (2016).

To ensure that the gaps highlighted in Figure 2 and Figure 3 represent real differences between subgroups rather than discrepant interpretations of items, the research team has conducted substantial measurement invariance testing. Using a differential item functioning (DIF) analysis, the team tested for differences in student performance by gender and by race at different grade levels. Overall, the team found only two items exhibiting moderate-to-large DIF where student subgroups in a few grades answered them differently (Meyer et al., 2018). Because only two significant findings out of such a large number of analyses could easily occur by chance, CORE can make a strong case that their scales engender consistent responding by different groups on specific items. However, it does not rule out that certain groups (e.g., African American, Latinx, or female students) answer the entire survey differently based on differences in interpretation or perspective. As discussed above, cultural or societal bias in survey response could result in students responding differently to every item within a scale, which would still represent a threat to validity. For example, if sixth-grade girls internalize a societal message that they should be humble, they would be responding to the same question differently than boys.

A related threat to validity can be found in ***reference bias***, and researchers have worked to rule out that students' responses may be influenced by peer norms (West, 2017). To do so, researchers compared the strength of the student-level correlations between social-emotional skills and academic indicators overall (i.e., across all students attending CORE middle schools) with those obtained when limiting the analysis to comparisons of students attending the same school. The logic of this exercise is straightforward: If students in higher performing schools rate themselves more critically, then average self-ratings in those schools will be artificially low.

This would bias the overall correlation downward, potentially making it lower than the correlations between student surveys within the same school environment. Figure 4 shows the result of this comparison for ELA test scores; the overall and within-school correlations differ modestly. However, the former are stronger than the latter—precisely the opposite pattern that would result from systematic reference bias due to varying expectations.

Figure 4. Student-Level Correlations Between Social-Emotional Skills and English Language Arts (ELA) Test Scores in CORE District Middle Schools



Source: West (2017).

Even given these positive results, much remains to be studied in determining how well the SEL measures fit the different contexts represented by all of the CORE schools. A particularly big challenge for CORE, and many other school districts across the country, is to devise optimal strategies for ensuring that students from different backgrounds are answering the surveys similarly. One important concern here is in ensuring that the survey is not biased against students in particular racial/ethnic groups. Educators in the CORE districts have chosen to measure SEL as a way to improve student outcomes, particularly for historically underserved students (Allbright et al., 2017; Marsh, McKibben, Hough, Hall, Allbright, Matewos, & Siqueira, 2018). However, some researchers are concerned that SEL frameworks may be biased against youth of color if they evaluate their behavior against a "white cultural frame of reference" (Gregory & Fergus, 2017). Further research is needed to unpack how students in different racial/ethnic groups are interpreting questions (i.e., through cognitive interviews), and to understand how cultural differences or a history of differential treatment in school (Okonofua, Walton, & Eberhardt, 2016; Tenenbaum & Ruck, 2007) may be related to observed SEL gaps.

Additionally, more work may be needed to better support students whose native language is not English. At present, forward and backward translation (e.g., having one person translate items from English to Spanish and then having a different person translate them back from Spanish to English so that these items can be compared against the originals) is generally accepted as the gold standard (Behling & Law, 2000). A parallel approach may be needed to ensure that students in different grades are interpreting items the same way. Especially if the goal is to track students' development over time, it will be important to disentangle real change in SEL from changes in cognitive process as students move from grade to grade. As CORE acquires more longitudinal data, it will be important to learn whether measures are interpreted in the same way by the same students over time, and to adjust for any differences in interpretation across time. For all of these questions about how students interpret survey questions, some form of cognitive interviewing could be helpful (Willis, 2005).

What Was the Level of Fidelity With Which CORE Acquired its Data?

For CORE, the ***survey administration*** process has been complicated by the need to allow for web-based surveys at some schools and paper-and-pencil surveys at other schools. Because of these modality issues, the different types of schools (elementary and secondary), variability in the size of schools, logistical differences in schedules, etc., it seems neither possible nor wise to attempt to standardize the administration procedures when administering an SEL survey at this scale. CORE dealt with these issues by hiring an external provider, Panorama Education, who navigated these administration challenges (West, Buckley, et al., 2017).

While it is up to every school and district specifically how they administer the survey, CORE developed guidance around survey administration to reduce some threats to validity, particularly those found in ***social desirability bias*** and to reduce the potential for some forms of measurement invariance (West, Buckley, et al., 2017). Specifically, student responses are confidential, so that their teachers cannot see how students respond. Additionally, the adults proctoring survey administration were asked to stand at the back of the classroom instead of circulating, and demographic questions were either not asked (if responses could be linked to administrative data) or asked at the end of the survey (to prevent the possibility that identifying with particular demographic groups might bias their subsequent answers).

More needs to be understood with respect to ascertaining how much effort students put into taking the surveys. One analysis suggests that the average item missing rates range from a little more than 1 percent to a little less than 7 percent per scale—with younger grades having much higher missing rates than their older peers (Meyer et al., 2018). It could be useful to contrast survey administration practices at schools with different rates of missing data and then test to see whether implementing practices and procedures from the schools with more successful administrations might improve data quality at other schools.

Similarly, it seems important to investigate the extent of ***satisficing*** on the SEL surveys—are students rushing or responding carelessly just to finish the task? Although it is harder to detect socially desirable responding, CORE's SEL scales do not include any obviously embarrassing questions, reducing the likelihood that social desirability bias is a big problem. On the other hand, the data in Figure 2 shows that girls' self-efficacy is slightly higher than boys in the early grades but may decline more rapidly than boys from sixth through ninth grades. Whether this finding reflects true changes in self-efficacy, changes in norms to respond modestly (which might be more socially desirable for girls than for boys), or both is unclear. If girls report what they think is expected of them (modesty) rather than what they actually believe, it could be a substantial source of error (Tourangeau, Rips, & Rasinski, 2000). Future research can improve our understanding of threats to validity found in survey administration by: detecting satisficers; better understanding the effects of fatigue, survey length, respondent age, and reading level on the fidelity of data; and learning from metadata (e.g., click-counts, timers, eye-gaze tracking data, etc.) about student engagement in the surveys.

To What Extent Are the CORE Data Being Used Appropriately?

In the area of ***analysis and measure creation***, the research team has taken multiple approaches to analyzing the CORE data and, as such, has a clear sense of the measurement properties of each scale (Meyer, Wang, and Rice, 2018). An important set of objectives here was to construct a measurement scale that: (1) provided robust measurement along a continuum of low to high "scores" on each of the constructs; (2) provided valid measurement of scores for all students, including those who may not have answered all of the survey questions; and (3) supported the option of modifying the set of survey questions over time to allow for continuous improvement in the quality and utility of the survey. The measurement team has applied scoring approaches using Item Response Theory (IRT) to develop IRT scale scores that help satisfy these objectives (Meyer et al., 2018). It is for this reason that researchers recommend using IRT scale scores instead of raw scores.[13]

Beyond establishing that the CORE SEL surveys have sound measurement properties, it is important to know that the resultant measures are related to other things we care about. This domain may represent where the CORE SEL surveys have the strongest evidence of validity. To begin, the CORE SEL scales show substantial evidence of ***convergent validity,*** meaning that they are correlated with other, related measures. For example, there is a strong relationship between students' reports on the social-emotional surveys and teacher, student, and staff reports about school culture and climate, connecting SEL to school-level practices that are hypothesized to improve it (Hough et al., 2017; Kraft, Buckley, Ruzek, Schenke, & Hulleman, 2018). Another study established that students' SEL reports are connected to indicators of persistence on computer adaptive tests (Soland, Jensen, Keys, Wolk, & Bi, Forthcoming). And yet another study has connected students' self-reports of their own SEL to teacher reports of the perceived SEL of each student (Scharer, West, & Dow, 2017).

Additionally, researchers working with the CORE data have invested a great deal of time in establishing ***predictive validity,*** showing that the school-level SEL indicators are correlated

[13] It is important to note that the CORE districts currently report aggregated raw scores rather than IRT scale scores for ease of interpretation and transparency. Student responses on the SEL surveys are translated into the percentage of positive responses in each school; for example, a school with a score of 80 would indicate that 80 percent of the survey questions were answered positively by students.

with students' grade point average, test scores (in math and English), suspension rates, and absence rates (West, 2017; Hough et al., 2017). Studies have also shown that high SEL reports, for example, on growth mindset, are related to growth in academic achievement at the student level (Claro & Loeb, 2017). Establishing that the SEL scales are related to other academic and behavioral outcomes is an important first step in establishing predictive validity; however, the next step is to test whether these associations are causal: Do interventions that change SEL cause changes in other student outcomes?

Along these lines, some of the most exciting work underway in this area is in understanding the extent to which these measures change over time at the student level, and how schools are impacting this student growth. If CORE's SEL measures are to be used to understand school performance, we need to know that students' development over time can be detected by the measures, and that schools can and do affect students' development. Existing evidence is promising that a school value-added model can be developed and utilized (Loeb, Christian, Hough, Meyer, Rice, & West, 2018). Researchers have shown that a school growth model can be constructed for the SEL surveys, that there is true variation in the extent to which schools contribute to student growth on these measures, and that the school effects on each construct are related to one another and to academic outcomes (although more weakly).

However, two important results provide evidence that SEL value-added models are not yet ready for practical application: (a) The goodness of the fit of the model is weaker in the SEL value-added models than in value-added models of academic outcomes, which raises questions about how well student growth is identified; and (b) there is not much across-school variation in the level of SEL outcomes, which is not what would be expected if there were persistent effects of schools on SEL outcomes. More research is needed to fully understand school effects on SEL; more years of data will allow the research team to investigate stability in effects over time, as well as to disentangle school effects from classroom or teacher effects, and to better understand the role of school and classroom context on students' social-emotional development.

Regardless of whether districts are using SEL surveys to establish a baseline, predict other outcomes, or measure change over time, they will want to attend to the ***consequential validity*** of how they use students' survey scores. Originally, the CORE districts planned to use these SEL measures as a part of an accountability policy that would examine the measures only when aggregated to the school level (Allbright et al., 2017). By the time the pilot testing was completed and the measures were ready to be used with stakes attached to them, No Child Left Behind had been replaced by the Every Student Succeeds Act, and the CORE districts were no longer required to use these scores as part of an accountability formula. While there were never strict consequences attached to the SEL measures in CORE, research has shown that their inclusion in a system of measures used for accountability could result in a very different set of schools being identified for improvement (Hough et al., 2017; Hough, Penner, & Witte, 2016). The consequences attached to the SEL survey measures could lead to some of the well-documented, negative, unintended consequences that can accompany measures used in accountability systems (Figlio & Getzler, 2002; Jacob & Levitt, 2003; Lauen & Gaddis, 2015; Neal & Schanzenbach, 2010). However, since the SEL surveys were never used in this way, the effect of accountability on SEL survey-taking remains unknown from the CORE experience.

However, measuring SEL to track school performance, even in the absence of the NCLB accountability requirement, sends a strong signal about what is valued in CORE districts and schools, pointing to the ***survey as an intervention***. Although they are not used in a formal accountability system, the surveys are used to track school performance, and in two districts results are even reported publicly (Marsh et al., 2018). Researchers have found that just the act of measuring SEL changed perceptions about what outcomes the schools felt they should be working toward. Educators in the CORE districts reported that measuring SEL gave them license to focus on it, and that thinking about school improvement in more than purely academic terms encouraged school leaders to broaden their conceptions of student and school success (Marsh et al., 2016). The consequences associated with these measures may have also led to CORE's unusually high response rates. For example, in 2014–15, CORE-wide, 74 percent of students in Grades 5–12 completed the SEL survey; the median elementary school had an 86 percent response rate; the median middle school, 84 percent; and the median high school, 75 percent (Hough et al., 2017). However, it is worth remembering that high response rates are typically much less important than obtaining a representative sample if a school's goal is to use the survey data as a basis for decision-making.

Within these considerations of appropriate data use, several tensions loom as particularly important areas for future research, debate, and innovation. First, schools usually strive to keep surveys short to minimize the loss of instructional time. Yet, measures become less reliable with fewer items (West, Pier, et al., 2018). Typically, schools can get away with shorter surveys if making decisions about large numbers of students, because the large sample of students provides enough signal to cut through the noise. Alternately, schools could switch from short surveys to extensive survey batteries that are administered repeatedly if making decisions at the student level, because the large sample of items could provide enough signal to compensate for the small number of students. However, at present, there are no clear avenues through which one can keep survey assessments short and still maintain high levels of accuracy for individual predictions. Relatedly, research shows that sophisticated scoring approaches yield more reliable results, yet educators may find them hard to interpret and less usable as compared to simpler statistics. These tensions, between measurement and use, are essential to navigate when considering the validity of these metrics.

Second, at a more philosophical level, many schools are eager to employ SEL surveys as a means to predicting students' achievement. However, many would argue that major SEL constructs—the social connectedness, motivation, and emotional well-being of our youth—are higher priorities than their academic attainment and should be treated as valued outcomes in and of themselves. For those who are eager to use SEL surveys for their predictive power, it is unclear what the right level of prediction would be. If SEL constructs were perfectly correlated with students' test scores, the measures would be redundant and only one would be needed. So a final lingering question for these measures is whether they are best thought of as important ends in their own right, as predictors of other student outcomes (begging questions about appropriate levels of prediction), or both.

Finally, when a new measurement system is introduced for a particular set of students in schools with a particular context, it is important to ask well the lessons learned will ***generalize*** beyond the context in which they were tested. Perhaps most helpful to CORE and to the field more broadly will be any techniques that can help predict the extent to which certain types of validity evidence (e.g., structural validity or reliability) will generalize from one population or setting to another.

Conclusion

Like any important educational policy decision, the question of whether a particular district should incorporate student perception surveys into its assessment system will depend upon a host of factors. Inevitably, smart decisions will depend on nuances of the context. Thus, in this brief, we explicitly avoided trying to make a blanket recommendation regarding the CORE surveys. Instead, we strove to provide a pragmatic framework to facilitate thinking through these complicated issues of validity.

In conclusion, we should underscore that the biggest choices a district might make are largely philosophical in nature. Districts might ask themselves: What outcomes do we value most (Brighouse, Ladd, Loeb, & Swift, 2016)? CORE examined self-efficacy but not teacher–student relationships—is that the right choice for our district? Should we prioritize growth mindset now, but shift towards examining students' valuation of the subject matter in the future?

Nearly as important is the question of how much evidence of validity there is for the measures under consideration. Many types of validity exist that one would want to accumulate over time to optimize one's confidence in a measure. CORE has strong evidence for validity in their measures of social awareness, self-management, self-efficacy, and growth mindset. However, as is always the case, the evidence could be stronger, and will continue to strengthen with time, as research continues and as the measures themselves are improved as a result (Davidson, Crowder, Gordon, Domitrovich, Brown, & Hayes, 2017). Because validity is a process, the coming years are likely to shed more light on the strengths and limitations of these measures.

In advocating for a pragmatic approach to validity, we presented four simple questions that cover the main domains for district leaders to think through with respect to whether the CORE survey scales or some other collection of social-emotional learning measures would be appropriate to implement in their district:

1. In evaluating ***how well the measures were designed***, district leaders can examine certain psychometric properties of the scales—how reliable are they and do they form a single coherent factor? However, they can also consult with experts to decide

whether the right content is included in each scale, whether a representative cross-section of indicators is present, and whether the items themselves adhere to best practices in survey design.

2. Determining ***whether the measures fit the local school context*** requires knowing whether the scales are sensitive enough to array students along a continuum (as opposed to suffering from floor or ceiling effects), are susceptible to reference bias, and/or are perceived differently by different subgroups of students (i.e., they have measurement invariance issues). District leaders will also want to keep in mind that CORE represents large, urban districts in California—school leaders across different locations will need to consider which contextual differences might affect the validity of the measures.
3. For many district leaders, how carefully the data were collected for CORE may be less of a concern than ***whether data could be collected with fidelity in their own district***—can strong survey administration practices that minimize satisficing and social desirability bias be employed?
4. Lessons from CORE on the ***extent to which data are being used appropriately*** may be particularly useful. The CORE districts went out of their way to ensure that the SEL data were used as "a flashlight, not a hammer," and built a system to ensure that what indicators reveal about school performance is used to help them improve rather than punish (Marsh et al., 2016). Using such measures to hold teachers, principals, or schools accountable could have dramatic impacts on their validity and could introduce unintended consequences. Furthermore, as other schools and districts consider using such measures, it is important the system leaders stress an improvement mindset, and focus on how schools and teachers can improve their practice to better serve students. New measures present new opportunities to understand how schools are serving diverse students, and can prompt educators and stakeholders to have honest conversations about how to develop inclusive, equitable school environments. However, without this systemic focus on improvement and equity, SEL measures could be misused to further scapegoat already vulnerable populations.

We have argued for the value of thinking of validity as an ongoing process. In the same way, putting in place an assessment system that works well for a particular district is a process. Mistakes will inevitably be made and hopefully important learnings can be extracted from those mistakes. CORE has put in place a strategic partnership with PACE to help ensure that all constituents are learning from this process. We view this as a healthy part of the process. In this way, we recommend that the introduction of new measures should include a clear mechanism for studying them to continually improve them, and their use, over time.

References

Allbright, T., Marsh, J., & Hough, H. J. (2017). *More than test scores: Designing accountability systems that include noncognitive factors.* Paper presented at the annual meeting of the American Education Research Association. San Antonio, TX.

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

Barge, S., & Gehlbach, H. (2012). Using the theory of satisficing to evaluate the quality of survey data. *Research in Higher Education*, *53*(2), 182–200. doi:10.1007/s11162-011-9251-2

Behling, O., & Law, K. S. (2000). *Translating questionnaires and other research instruments: problems and solutions*. Thousand Oaks, CA: Sage Publications.

Benson, J., & Hocevar, D. (1985). The impact of item phrasing on the validity of attitudes scales for elementary school children. *Journal of Educational Measurement*, *22*(3), 231–240. doi:10.1111/j.1745-3984.1985.tb01061.x

Blad, E. (2017, October 4). No state will measure social-emotional learning under ESSA: Will that slow its momentum? *Education Week.*

Bolt, D. M., Wang, Y. C., Meyer, R. H., & Rice, A. B. (2018). *IRT mixture model for rating scale confusion associated with negatively worded items.* Paper presented at the National Council on Measurement in Education annual conference, New York, NY.

Brighouse, H., Ladd, H. F., Loeb, S., & Swift, A. (2016). Educational goods and values: A framework for decision makers. *Theory and Research in Education*, *14*(1), 3–25.

Claro, S., & Loeb, S. (2017). *Effect of Growth Mindset on Achievement: Evidence from California CORE School Districts.* Paper presented at the Annual Meeting of the Association for Public Policy Analysis and Management. Chicago, IL.

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.

Davidson, L. A., Crowder, M. K., Gordon, R. A., Domitrovich, C. E., Brown, R. D., & Hayes, B. I. (2017). A continuous improvement approach to social and emotional competency measurement. *Journal of Applied Developmental Psychology*.

DeVellis, R. F. (2003). *Scale development: Theory and applications* (2nd ed.). Newbury Park, CA: Sage.

Dillman, D. A., Smyth, J. D., & Christian, L. M. (2014). *Internet, phone, mail, and mixed-mode surveys: The tailored design method* (4th ed.). Hoboken, NJ: John Wiley & Sons.

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.

Dusenbury, L., Dermody, C., & Weissberg, R. P. (2018). *2018 State scorecard scan*. CASEL. Retrieved from https://casel.org/wp-content/uploads/2018/02/2018-State-Scan-FINAL.pdf

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

Figlio, D. N., & Getzler, L. S. (2002). Accountability, ability and disability: Gaming the system. Retrieved from http://www.nber.org/papers/w9307

Fowler, F. J. (2009). *Survey research methods* (4th ed.). Thousand Oaks, CA: Sage Publications.

Gehlbach, H. (2015). Seven survey sins. *The Journal of Early Adolescence*, 35, 883–897. doi:10.1177/0272431615578276

Gehlbach, H., & Artino, A. R. (2018). The survey checklist (manifesto). *Academic Medicine: Journal Of The Association Of American Medical Colleges*, *93*(3), 360–366. doi:10.1097/ACM.0000000000002083

Gehlbach, H., & Barge, S. (2012). Anchoring and adjusting in questionnaire responses. *Basic and Applied Social Psychology*, *34*(5), 417–433. doi:10.1080/01973533.2012.711691

Gehlbach, H., & Brinkworth, M. E. (2011). Measure twice, cut down error: A process for enhancing the validity of survey scales. *Review of General Psychology*, *15*(4), 380–387. doi:10.1037/a0025704

Gehlbach, H., Robinson, C. D., Finefter-Rosenbluh, I., Benshoof, C., & Schneider, J. (2018). Questionnaires as interventions: Can taking a survey increase teachers' openness to student feedback surveys? *Educational Psychology*, *38*(3), 350–367. doi:10.1080/01443410.2017.1349876

Gregory, A., & Fergus, E. (2017). Social and Emotional Learning and Equity in School Discipline. In S. M. Jones & E. Doolittle (Co-Eds.), *Social-emotional learning*, *27*(1). Princeton. Retrieved from https://futureofchildren.princeton.edu/file/986/download?token=WkE8Dw_D

Guzman-Lopez, A. (2017, April 10). *California schools innovating how to measure social-emotional learning.* Retrieved from https://www.scpr.org/news/2017/04/10/70606/california-schools-innovating-how-to-measure-socia/

Hough, H. J., Kalogrides, D., & Loeb, S. (2017). *Using surveys of students' social-emotional skills and school climate for accountability and continuous improvement.* Stanford, CA: Policy Analysis for California Education (PACE). Retrieved from https://edpolicyinca.org/publications/using-sel-and-cc

Hough, H. J., Penner, E., & Witte, J. (2016). *Identity crisis: Multiple measures and the identification of schools under ESSA.* Stanford, CA: Policy Analysis for California Education (PACE). Retrieved from http://www.edpolicyinca.org/publications/identity-crisis-multiple-measures-and-identification-schools-under-essa

Jacob, B. A., & Levitt, S. D. (2003). *Rotten apples: An investigation of the prevalence and predictors of teacher cheating.* Retrieved from http://www.nber.org/papers/w9413

Kane, M. T. (1992). An argument-based approach to validity. *Psychological Bulletin*, *112*(3), 527–535. doi:10.1037/0033-2909.112.3.527

Kane, M. T. (2006). Validation. In R. L. Brennan (Ed.), *Educational Measurement* (4 ed., pp. 17–64). Westport, CT: Praeger.

Kenny, D. A. (1995). The multitrait-multimethod matrix: Design, analysis, and conceptual issues. In P. E. Shrout & S. T. Fiske (Eds.), *Personality research, methods, and theory: A festschrift honoring Donald W Fiske* (pp. 111–124). Hillsdale, NJ, England: Lawrence Erlbaum Associates.

Krachman, S. B., Arnold, R., & LaRocca, R. (2016). *Expanding our definition of student success: A case study of the CORE districts*. Boston, MA: Transforming Education.

Kraft, M., Buckley, K., Ruzek, K., Schenke, K., Hulleman, C. (2018). *The effect of school climate on students' social-emotional competencies.* Paper presented at the annual meeting of the American Education Research Association. New York, NY.

Krosnick, J. A. (1991). Response strategies for coping with the cognitive demands of attitude measures in surveys. *Applied Cognitive Psychology*, *5*(3), 213–236.

Krosnick, J. A., & Presser, S. (2010). Question and questionnaire design. In P. V. Marsden & J. D. Wright (Eds.), *Handbook of Survey Research* (2nd ed.). Bingley, UK: Emerald Group Publishing.

LaRocca, R., & Krachman, S., (2017). *Expanding the definition of student success under ESSA opportunities to advance social-emotional mindsets, skills, and habits for today's students.* Boston, MA: Transforming Education.

Lauen, D. L., & Gaddis, S. M. (2015). Accountability pressure, academic standards, and educational triage. *Educational Evaluation and Policy Analysis*. doi:0162373715598577

Loeb, S., Christian, M., Hough, H., Meyer, R., Rice, A., & West, M. (2018). *School effects on social-emotional learning: Findings from the first large-scale panel survey of students*. Stanford, CA: Policy Analysis for California Education (PACE). Retrieved from www.edpolicyinca.org/publications/sel-school-effects

Marsh, J. A., Bush-Mecenas, S., Hough, H. J., Park, V., Allbright, T., Hall, M., & Glover, H. (2016). *At the forefront of the new accountability era: Early implementation findings from the CORE waiver districts*. Stanford, CA: Policy Analysis for California Education (PACE). Retrieved from https://edpolicyinca.org/publications/at-the-forefront

Marsh, J. A., McKibben, S., Hough, H. J., Hall, M., Allbright, T., Matewos, A., & Siqueira, C. (2018). *Enacting social-emotional learning: Practices and supports employed in CORE districts and schools.* Stanford, CA: Policy Analysis for California Education (PACE). Retrieved from http://www.edpolicyinca.org/publications/sel-practices

McCormick, M. P., Cappella, E., O'Connor, E. E., & McClowry, S. G. (2015). Context matters for social-emotional learning: Examining variation in program impact by dimensions of school climate. *American Journal of Community Psychology*, *56*(1–2), 101–119.

Messick, S. (1995). Validity of psychological assessment: Validation of inferences from persons' responses and performances as scientific inquiry into score meaning. *American Psychologist*, *50*(9), 741–749. doi:10.1037/0003-066X.50.9.741

Meyer, R., Wang, C., & Rice, A. (2018). *Measuring Students' Social-Emotional Learning Among California's CORE Districts: An IRT Modelling Approach.* Stanford, CA: Policy Analysis for California Education (PACE). Retrieved from http://www.edpolicyinca.org/publications/sel-measurement

Nagaoka, J., Farrington, C. A., Ehrlich, S. B., & Heath, R. D. (2015). *Foundations for young adult success: A developmental framework*. Concept Paper for Research and Practice. University of Chicago Consortium on Chicago School Research.

Nayfack, M., Park, V., Hough, H., & Willis, L. (2017). *Building systems knowledge for continuous improvement: Early lessons from the CORE districts.* Stanford, CA: Policy Analysis for California Education (PACE). Retrieved from www.edpolicyinca.org/publications/building-systems-knowledge-for-continuous-improvement

Neal, D., & Schanzenbach, D. W. (2010). Left behind by design: Proficiency counts and test-based accountability. *The Review of Economics and Statistics*, *92*(2), 263–283.

Okonofua, J. A., Walton, G. M., & Eberhardt, J. L. (2016). A vicious cycle: A social–psychological account of extreme racial disparities in school discipline. *Perspectives on Psychological Science*, *11*(3).

Robinson-Cimpian, J. (2014). Inaccurate estimation of disparities due to mischievous responders: Several suggestions to assess conclusions. *Educational Researcher*, *43*(4), 171–185.

Rubio, D. M., Berg-Weger, M., Tebb, S. S., Lee, E. S., & Rauch, S. (2003). Objectifying content validity: Conducting a content validity study in social work research. *Social Work Research*, *27*(2), 94–104.

Scharer, E., West, M., & Dow, A. (2017). *A teacher's perspective on students' socio-emotional skills: Can teacher assessments improve our understanding of student social-emotional skills?* Paper presented at the annual meeting of the Association for Education Finance and Policy, Washington, DC.

Schweig, J., Hamilton, L. S., Stecher, B. M., & Baker, G. (2017). *Building a repository of social and emotional learning assessments.* Paper presented at the Annual Meeting of the Association for Public Policy Analysis and Management. Chicago, IL.

Soland, J., Jensen, N., Keys, T., Wolk, E., & Bi, S. (Under review). *Is low test motivation a sign of disengagement from school? Examining indicators of dropout conditional on rapid guessing scores.* Educational Assessment.

Tenenbaum, H. R., & Ruck, M. D. (2007). Are teachers' expectations different for racial minority than for European American students? A meta-analysis. *Journal of Educational Psychology*, *99*(2).

Tourangeau, R., Rips, L. J., & Rasinski, K. A. (2000). *The psychology of survey response*. New York: Cambridge University Press.

Tourangeau, R., & Yan, T. (2007). Sensitive questions in surveys. *Psychological Bulletin*, *133*(5), 859–883.

West, M. (2017). Should non-cognitive skills be included in school accountability systems? Preliminary evidence from California's CORE districts. *Evidence Speaks Reports*, *1*(13). Brookings Institute.

West, M. R., Buckley, K., Krachman, S. B., & Bookman, N. (2017). Development and implementation of student social-emotional surveys in the CORE districts. *Journal of Applied Developmental Psychology*.

West, M. R., Kraft, M. A., Finn, A. S., Martin, R. E., 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. doi:10.3102/0162373715597298

West, M., Pier, L., Fricke, H., Hough, H. J., Loeb, S., Meyer, R., & Rice, A. (2018). *Measuring and charting the development of student social-emotional learning: Evidence from the first large-scale panel survey of students.* Stanford, CA: Policy Analysis for California Education (PACE). Retrieved from https://edpolicyinca.org/publications/sel-trends

Willis, G. B. (2005). *Cognitive interviewing: A tool for improving questionnaire design*. Thousand Oaks, CA: Sage Publications.

Zernike, K. (2016, February 29). Testing for joy and grit: Schools nationwide push to measure students' emotional skills. *The New York Times*. Retrieved from http://www.nytimes.com/2016/03/01/us/testing-for-joy-and-grit-schools-nationwide-push-to-measure-students-emotional-skills.html

About the Authors

Hunter Gehlbach is an Associate Professor at the University of California, Santa Barbara Gevirtz Graduate School of Education and the Director of Research at Panorama Education. As Gehlbach is an educational psychologist by training and a social psychologist at heart, his primary interests lie in improving the social climate of schools, bolstering students' motivation, and helping adolescents develop greater self-regulatory capacities. Complementing these substantive interests, he helps social scientists and practitioners design better questionnaires. As Director of Research at Panorama, Gehlbach has applied his knowledge of survey design and administration across a diverse array of schools and districts. A former tenth-grade social studies teacher, Hunter holds degrees from Swarthmore College (B.A.), the University of Massachusetts-Amherst (M.Ed. in school counseling), Stanford (M.A. in social psychology; Ph.D. in educational psychology), and completed a post-doctoral fellowship at the University of Connecticut. He was on the faculty at Harvard's Graduate School of Education from 2006–2015, before joining the faculty at UCSB. Gehlbach is a member of the questionnaire committees for the National Assessment of Educational Progress and the Programme for International Student Assessment.

Heather J. Hough is the Executive Director of the research partnership between Policy Analysis for California Education (PACE) and the CORE Districts, a collaborative of eight California school districts that have developed a robust measurement and accountability system representing over one million students. Before joining PACE, Heather was an improvement adviser with the Carnegie Foundation for the Advancement of Teaching, helping education system leaders use research and data to support continuous improvement. She has worked as a researcher with the Public Policy Institute of California, the Center for Education Policy Analysis at Stanford University, and the Center for Education Policy at SRI International. Heather's area of expertise is in district- and state-level policymaking and implementation, with a particular focus on policy coherence, system improvement, and school and teacher accountability. She holds a PhD in education policy and a B.A. in public policy from Stanford University.

Acknowledgements

We would like to thank the many individuals who contributed to this report. First, we are grateful to the generous sponsor of this brief, the Walton Family Foundation, in addition to the many funders who supported the research undergirding this summary, including the S. D. Bechtel, Jr. Foundation. We also thank all of the leaders and administrators in the CORE districts for their support throughout this project, along with the many school leaders and educators who participated in the research activities and shared their valuable time and insights with us. This project would not have been completed without the assistance of many colleagues at Education Analytics, Harvard University, Transforming Education, and Stanford University. Finally, we gratefully acknowledge the thoughtful feedback of two anonymous reviewers and multiple colleagues who reviewed earlier versions of this brief.

About the CORE-PACE Research Partnership

In October 2015, Policy Analysis for California Education (PACE) and the CORE Districts launched the CORE-PACE Research Partnership. This research partnership is focused on producing research that informs continuous improvement in the CORE districts and policy and practice in California and beyond. The CORE districts (Fresno, Garden Grove, Long Beach, Los Angeles, Oakland, Sacramento City, San Francisco, and Santa Ana Unified School Districts) together serve nearly one million students and utilize a unique multiple-measures data system to work together to improve student outcomes. Our research aims to deepen their learning, while sharing lessons more broadly to accelerate improvement across the state.

PACE
Policy Analysis for California Education

Stanford Graduate School of Education
520 Galvez Mall, CERAS 401
Stanford, CA 94305-3001
Phone: (650) 724-2832
Fax: (650) 723-9931

edpolicyinca.org