PACE
Policy Analysis for California Education
CORE-PACE RESEARCH PARTNERSHIP

WORKING PAPER

Measuring Students' Social-Emotional Learning Among California's CORE Districts: An IRT Modeling Approach

Robert H. Meyer
Education Analytics

Caroline Wang
Education Analytics

Andrew B. Rice
Education Analytics

With an increased appreciation of students' social-emotional skills among researchers and policy makers, many states and school districts are moving toward a systematic process to measure Social-Emotional Learning (SEL). In this study, we examine the measurement properties of California's CORE Districts' SEL survey administered to over 400,000 students in grades 3 to 12 during the 2015-16 school year. We conduct analyses through both classical test theory and item response theory frameworks, applying three different polytomous IRT models on both the full student sample and on separate samples from each grade. From these analyses, we summarize the psychometric properties of items at each grade level, compare items' functionality across grades, compare student outcomes from IRT models and the classical approach, make suggestions on approaches to modeling and scaling the SEL survey data, and identify items, by grade, that do not contribute positively to measurement of each outcome. Finally, we discuss policy implications in using SEL measures among educators, administrators, policy makers, and other stakeholders.

VERSION: May 2018

Acknowledgements
This paper was produced as part of the CORE-PACE Research Partnership, which is focused on producing research that informs continuous improvement in the CORE districts (Fresno, Garden Grove, Long Beach, Los Angeles, Oakland, Sacramento City, San Francisco, and Santa Ana unified school districts) and policy and practice in California and beyond. We thank the leaders and administrators in the CORE districts for their support throughout this project, and the generous funder of this research, the Walton Family Foundation. PACE working papers are circulated for discussion and comment purposes and have not undergone the peer-review process that accompanies official PACE publications.

Introduction

Research continually demonstrates the value of students' social-emotional skills, or noncognitive skills, such as growth mindset and self-management, in determining their future success, including academic achievement, workforce performance, and well-being (Cunningham & Villaseñor, 2016; de Ridder, Lensvelt-Mulders, Finkenauer, Stok, & Baumeister, 2012; Jones, Greenberg, & Crowley, 2015; Moffitt et al., 2011). For example, important social and emotional factors are shown in meta-analyses to promote success in school and life (Durlak, Weissberg, Dymnicki, Taylor, & Schellinger, 2011; Poropat, 2009; Taylor, Oberle, Durlak, & Weissberg, 2017). Their predictive power exceeds that of cognitive skills after controlling for educational attainment (Heckman, Humphries, & Kautz, 2014; Segal, 2013). Furthermore, teachers and school leaders have expressed their perceived value of SEL from practitioners' perspective; findings from a 2013 teacher survey with a nationally representative sample demonstrated that nearly all (93%) teachers believe that "SEL is very or fairly important for the in-school student experience" (Bridgeland, Bruce, & Hariharan, 2013, p.5). In addition, social and emotional skills in childhood predict higher long-term earnings and better financial situations in adulthood (Chetty et al., 2011; Moffitt et al., 2011). Surveys in the workplace show that employers value social and emotional skills as important for success; yet, this is where the greatest skill gaps exist (Cunningham & Villaseñor, 2016). Several longitudinal studies have also found statistically significant associations between measures of social-emotional skills and key young adult outcomes, across multiple domains in education, criminal activity, substance use, and mental health (Hawkins, Kosterman, Catalano, Hill, & Abbott, 2008; Jones, Greenberg, & Crowley, 2015).

More importantly, these social and emotional skills appear to be malleable during the early years; they can be substantially improved through early childhood interventions (Almlund, Duckworth, Heckman, & Kautz, 2011; Heckman, 2008). Research shows that schools can promote students' development of SEL through implementations of various policies and practices (Battistich, Schaps, & Wilson, 2004; Durlak, Weissberg, Dymnicki, Taylor, & Schellinger, 2011). Improving SEL has also been shown to be cost effective. A recent cost-benefit analysis of six SEL intervention studies showed a "substantial economic return:" For every $1 invested, there is a return of $11 (Belfield, Bowden, Klapp, Levin, Shand, & Zander, 2015, p.5).

However, there appears to be a gap between the existing research and teaching practices inside the classrooms. Despite the widespread recognition of the importance of SEL among teachers, parents, employers, and researchers, SEL is still considered "the missing piece in the educational puzzle" (Bridgeland, Bruce, & Hariharan, 2013, p.12). Moreover, there is only limited availability of large-scale surveys measuring personal characteristics. The relatively sparse research on development and measurement of SEL for educational purposes often has taken place at a small scale and with convenient samples of students, classrooms, and schools. Consequently, the measurement properties of these instruments and results likely imply limited generalizability.

Therefore, it is exciting to see recent policy changes at the federal and the state levels on SEL that may spur additional empirical research into SEL. The Every Student Succeeds Act (ESSA) of 2015 requires state accountability systems to include at least one indicator of school quality or student success other than students' cognitive abilities; this requirement signals a federal shift to a more holistic view of education that includes SEL factors (Elementary and Secondary Education Act of 1965, 2015, §1111). At the state level, all 50 states have integrated some degree of social and emotional content into their learning standards; many have legislative bills and policies in place to support statewide SEL implementation from preschool through 12th grade (Dusenbury, Newman, Weissberg, Goren, Domitrovich, & Mart, 2015). In addition, some districts are also moving toward a systematic process to measure students' SEL learning and incorporate SEL into applied settings (Oakland Unified School District, 2015).

California's CORE Districts are at the forefront of the national SEL movement. The CORE Districts are a consortium of eight California school districts—Fresno, Garden Grove, Long Beach, Los Angeles, Oakland, Sacramento, San Francisco, and Santa Ana—that collectively serve more than one million students attending roughly 1,800 schools. In 2013, CORE received a waiver from the U.S. Department of Education that allowed its member districts to waive key requirements of the No Child Left Behind school accountability system. Under this waiver, CORE sought to implement an accountability system that incorporated school performance across a broader range of outcome measures, rather than looking solely at standardized test scores and graduation rates. CORE's measurement system focuses on non-academic measures such as SEL and school culture/climate, alongside traditional academic indicators, to inform a more holistic index of school quality. Since then, CORE has transitioned to a support network focusing on continuous improvement based on the data it collects from the districts, and now seeks to leverage its SEL work in this mission.

CORE's SEL measures are generated from surveys of students in grades 3-12. Using three key criteria—meaningful, measurable, and actionable—CORE Districts prioritized four SEL skills, including self-management, growth mindset, self-efficacy, and social awareness (West, Buckley, Krachman, & Bookman, 2018). Using behavior rating scales, the SEL survey asks students to self-report their responses to each survey question on a 5-point Likert scale. Since the initial pilot in Spring 2014, around 445,000 students in 2014-15 and 484,000 students in 2015-16 in grades 3 through 12 have participated the SEL survey.

The process of measuring student's personal characteristics has raised many concerns; some known issues include student self-report and missing responses. Since its first administration, the measurement properties of the CORE SEL survey have not been extensively studied. Additionally, with the limited availability of large-scale surveys measuring social-emotional skills, no consensus exists on how best to score and report these surveys' outcomes.

This paper aims to investigate the measurement properties of the SEL items and to identify the best approach to model and score CORE's SEL survey. CORE's SEL survey is also unique because the same items are administered to a wide range of students in different grades. Therefore, we are also striving to understand whether students from different grades

perceive the same items differently via the insights provided by item response theory models. The following research questions motivated this research:

- Did the survey produce good measures of social-emotional skills?
- Did the items in each construct measure the same construct as intended?
- Did different types of students vary in how they responded to the items?
- Do survey results have the same meaning at different grade levels?
- Did the survey provide valid, reliable, and useful data for all students?
- How should the SEL survey be scored?

The next section of this paper describes the CORE Districts' SEL survey measures, followed by data used in this research. Under the framework of classical test theory (CTT) and item response theory (IRT), several measurement tools were used to address these questions. These psychometric analyses and their results are presented next. The final section provides a summary of the results and discussion of our findings.

CORE's SEL Measures

The CORE Districts' SEL survey is a suite of instruments designed to measure four SEL constructs: self-management (nine items), growth mindset (four items), self-efficacy (four items), and social awareness (eight items). Students in grades 3 through 12 rate themselves on the same 25 questions using a 5-point Likert scale. The four SEL constructs are designed to measure the following:

- **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).

It is worth mentioning that most survey items are positively phrased, but all four items on the growth mindset scale are negatively phrased. In the two example items listed below, the

first item is from the self-efficacy construct (phrased positively) and the second item is from the growth mindset construct (phrased negatively).

1. How confident are you about the following at school?
   I can earn an A in my classes.
   Not At All Confident (1), A Little Confident (2), Somewhat Confident (3), Mostly Confident (4), Completely Confident (5)

2. Please indicate how true each of the following statements is for you:
   My intelligence is something that I can't change very much.
   Not At All True (1), A Little True (2), Somewhat True (3), Mostly True (4), Completely True (5)

Research Sample

This paper focuses on CORE's SEL survey responses from the 2015-16 school year when seven districts—around 481,000 students from more than 1,200 schools—participated in the survey administration. Our research sample was selected based on two criteria. First, the SEL survey was originally written in English and translated into several other languages; only a small portion of the students took the survey in a language other than English, but the research included only those students who took the SEL survey in the original English version (note that we plan to compare the properties of items in English and other languages in future work). Second, like all survey data, not all students completed all items on the SEL survey. Our research sample included students who completed at least 50% of the items within each SEL construct. For example, the self-management construct has nine items in total. While conducting analysis of self-management, our final sample included students who answered five or more self-management items. If a student answered five self-management items but skipped all growth mindset items, this student was included in analyses related to self-management but was excluded in analyses related to growth mindset. Table 1 summarizes the number of students and schools at each grade in the research sample in the 2015-16 school year.

Table 1. Summary of Research Sample: Number of Students and Schools in Each Grade (2015–16)

	Number of students	Number of schools
Grade 3	38,644	495
Grade 4	56,181	722
Grade 5	55,437	727
Grade 6	44,136	342
Grade 7	43,032	198
Grade 8	43,023	198
Grade 9	40,723	222
Grade 10	39,671	233
Grade 11	34,439	225
Grade 12	32,405	225
Total	427,691	1,091

A subset of the research sample was used in the differential item functioning (DIF) analysis. Given that the number of items within each construct can be as few as four, we included students who completed all items on the studied construct in the DIF analysis to minimize the complications of missing responses and increase the estimation precision. Table 2 summarizes the percentage of students in different demographic groups (e.g., gender, race/ethnicity) used in the DIF sample in 2015-16. As shown in Table 2, a majority of the students were Hispanic/Latino (70.1%). Additionally, a large percentage of students participating in the SEL survey were socioeconomically disadvantaged students (78.2%).

Table 2. Description of Student Sample Components in 2015–16

Demographic Background	Percentage
Male	50.6%
Female	49.4%
White	8.3%
African American	8.1%
Hispanic/Latino	70.1%
Asian	7.5%
Other race/ethnicity	6.0%
Socioeconomically disadvantaged students	78.2%
English language learners	19.4%
Students with disabilities	11.0%

Analyses

In order to examine the psychometric properties of the SEL instruments, we conducted the following series of analyses under the framework of CTT and IRT. The analyses were conducted separately for each of the four SEL constructs – self-management, growth mindset, self-efficacy, and social awareness – and at each grade level from grade 3 to grade 12.

Classical Item Analysis

This includes analyses under the framework of CTT. For each SEL construct, we examined the internal consistency as measured by Cronbach's alpha. For each item, we calculated item difficulty (i.e., p-value), item discrimination (i.e., polyserial correlation), and percent missing within each domain at each grade level. These analyses can help evaluate the overall properties of the scales, detect whether there are individual items that contribute more or less to the measurement of the constructs (e.g., Meyer, Gawade, & Wang, 2016; Wang, Gawade, & Meyer, 2015), and determine if/how these properties change across grades and various student subgroups.

Factor Analysis

Since the SEL survey is designed with four domains, we conducted exploratory factor analysis to examine the factor structure of the survey items. Results from factor analysis are helpful in assessing whether and which items cross load onto more than one construct. They are also helpful in determining whether to use unidimensional or multidimensional IRT models.

Differential Item Functioning Analysis (DIF)

DIF statistics indicate whether there are statistically significant differences in student performance on an item between the reference group (e.g., male) and the focal group (e.g., female) who performed similarly on the entire survey. We conducted DIF analysis using the standardized mean difference method (SMD) in conjunction with the Mantel chi-square statistic (Mantel, 1963; Mantel & Haenszel, 1959). The SMD method was originally developed by Dorans and Kulick (1986) to estimate DIF for dichotomous items. Dorans and Schmitt (1991) extended this statistic to estimate DIF in the case of polytomous items. This extension was used in this research to handle polytomously scored SEL survey items. It can be calculated using the following formula:

$$SMD = \left(\sum_{k=1}^{K} \frac{n_{F+k}}{n_{F++}} \frac{\sum_{j=1}^{J} y_j n_{Fjk}}{n_{F+k}} \right) - \left(\sum_{k=1}^{K} \frac{n_{F+k}}{n_{F++}} \frac{\sum_{j=1}^{J} y_j n_{Rjk}}{n_{R+k}} \right) \quad (1)$$

where *K* is the number of score categories in the total score; *J* is the number of score categories in the studied item; y_j is the score of the studied item at the j^{th} score category; and the variable n_{ijk} represents the number of examinees in group *i* (*F* = focal or *R* = reference) who obtained a

score of y_j on the studied item and a total score of x_k. The total score, or the criterion score, is calculated as the mean of all items in the corresponding construct since missing responses were excluded from DIF analysis.

In addition, the Mantel χ^2statistic was used to test the null hypothesis that there is no DIF in the studied item, where χ^2 is distributed as a chi-square variable with 1 degree of freedom. Large values of Mantel chi-square statistic with *p*-values smaller than 0.05 provide evidence that the item exhibits DIF.

Based on the *SMD* and Mantel χ^2 statistics, items can be classified into three categories (see Table 3). We applied DIF flagging criteria of these categories, which are used widely by assessment programs such as NAEP, PARCC, and SBAC (Allen, Donoghue, & Schoeps, 2001; Pearson, 2017; SBAC, 2016), to this study. Specifically, category A items exhibit negligible DIF, category B items exhibit slight-to-moderate DIF, and category C items exhibit moderate-to-large DIF. Moreover, a positive *SMD* statistic means that the focal group has a higher item mean score than the reference group, conditioning on the construct mean score. In contrast, a negative *SMD* statistic means that the focal group has a lower item mean score than the reference group, conditioning on the construct mean score.

Table 3. DIF Categories for CORE's SEL Survey Items

DIF Category	Criteria
A (negligible)	Mantel Chi-square *p*-value > 0.05 or $\lvert SMD/SD \rvert \leq 0.17$
B (slight to moderate)	Mantel Chi-square *p*-value < 0.05 and $\lvert SMD/SD \rvert > 0.17$
C (moderate to large)	Mantel Chi-square *p*-value < 0.05 and $\lvert SMD/SD \rvert > 0.25$

Note: *SMD* = standardized DIF calculated under equation (1); *SD* = total group standard deviation of item score.

We analyzed DIF among three types of demographic subgroups: gender, racial/ethnic subgroups, and English language learners and their peers. These results can provide evidence on whether students from different demographic groups interpret the same item similarly. Also, DIF analysis can help in identifying unintended factors that interfere with measurement, and thus enable further understanding of the measurement properties of the survey instruments and improve their generalizability across sub-populations.

Item Response Theory Modeling

We applied IRT models to better understand item-level psychometric properties of the SEL survey and identify potential problematic issues in the items. We utilized an IRT calibration/scaling model and estimated scale scores from grades 3 through 12 for each SEL construct. IRT-based item and test information can provide insights into how individual items contribute to each construct, which can help guide us through the survey refinement stage. IRT can also function as a scoring model, in that it can accommodate missing responses and take both item characteristics and response patterns into consideration. IRT also provides a basis for defining comparable scores across grades and across years. Finally, IRT models can provide insight into item interpretation issues, by offering the opportunity to learn more about how

students at various age groups respond to items differently, and therefore may perceive the same item differently.

Due to limited large-scale survey practice, there is currently no consensus on which IRT model one should use when modeling survey data. The Programme for International Student Assessment (PISA) administered by the Organisation for Economic Co-operation and Development (OECD) has used both the partial credit model (PCM; Masters, 1982) and the generalized partial credit model (GPCM; Muraki, 1992) on PISA's background questionnaires (OECD, 2017). Other large-scale international comparison studies, such as the Trends in International Mathematics and Science Study (TIMSS) and the International Civic and Citizenship Education Study (ICCS) have been using PCM to form their context questionnaire scales (Martin, Mullis, Hooper, Yin, Foy, & Palazzo, 2016; Schulz, Ainley, & Fraillon, 2011). In addition, the nominal response model (NRM; Bock, 1972) has also been used in research to identify unexpected behavior in responses of polytomous items. For example, Sharp et al. (2012) used NRM to detect items whose responses were not in the order as expected and items whose response categories should be collapsed (e.g., the middle two responses in a four-category item, "never" (1), "sometimes" (2), "most of the time" (3), and "almost always" (4), should be collapsed to yield a three-category item). Preston, Reise, Cai, and Hays (2011) proposed to use category boundary discrimination parameters derived from the nominal model to help with "testing the assumption of ordered response categories inherent in most psychological scales…, identify poorly functioning response categories…, [and] studying the effects of reverse wording" (p.548).

In this study, we explored the application of three unidimensional polytomous IRT models to the data—the PCM, the GPCM, and the NRM, each of which makes varying assumptions about the data that can be empirically evaluated. The R package mirt (version 1.25) was used in IRT modeling.

The PCM is an extension of the Rasch model allowing it to handle polytomous response data. Its mathematical form can be written as

$$P_{ix}(\theta) = \frac{exp \sum_{k=0}^{x}(D(\theta - b_{ik}))}{\sum_{h}^{m_i} exp \sum_{k=0}^{h}(D(\theta - b_{ik}))}$$

$$x = 0,\ 1,\ \ \ldots,\ m_i \qquad (2)$$

where $P_{ix}(\theta)$ is the probability of a person with ability θ scoring x on item i; item i has m_i "steps" and $m_i + 1$ score values ranging from 0 to m_i; b_{ik} is the difficulty of step k of item i; and D is a scaling factor set to 1.7 to approximate the normal ogive model.

The GPCM is a generalization of the PCM by adding an item discrimination parameter to the model. It can be expressed as

$$P_{ix}(\theta) = \frac{exp\sum_{k=0}^{x}(Da_i(\theta - b_i + d_{ix}))}{\sum_{h}^{m_i} exp\sum_{k=0}^{h}(Da_i(\theta - b_i + d_{ix}))}$$

$$x = 0,\ 1,\ \ \ldots,\ m_i \tag{3}$$

where a_i is an item discrimination parameter or slope parameter; b_i is an item location parameter; and d_{ix} is a category parameter. With $m_i + 1$ categories, only m_i category parameters can be identified.

Unlike PCM and GPCM introduced above, where the $m_i + 1$ polytomous item responses from 0 to m_i are in order, polytomous responses in NRM are not assumed to be in order. In the case of the SEL survey, although item responses are coded numerically from 0 to 4, the values of these responses do not represent the actual values of the scores, but nominal indications for response categories. The NRM can be written as

$$P_{ix}(\theta) = \frac{exp(a_{ix}\theta + c_{ix})}{\sum_{0}^{m_i} exp(a_{ix}\theta + c_{ix})}$$

$$x = 0,\ 1,\ \ \ldots,\ m_i \tag{4}$$

where a_{ix} is a slope parameter; c_{ix} is an intercept parameter. With $m_i + 1$ categories, a set of $m_i + 1$ slop and intercept parameters are estimated for each response category for an item.

We estimated these three polytomous IRT models for each SEL construct and each grade level. With 10 grade levels (from grade 3 to grade 12), four SEL constructs (self-management, growth mindset, self-efficacy, and social awareness), and three IRT models, a total of 120 separate IRT item calibrations were conducted. At the end of each calibration, we examined model-data fit, model convergent status, item parameter estimates and standard errors, and item category response function plots. We compared the fit of these models to determine which is most useful for our applications. We compared item functionality across grade levels. We also compared student outcomes estimated using different polytomous IRT models as well as the classical approach (i.e., raw mean scores excluding missing responses). In comparison to IRT scores and CTT raw scores, IRT true scores were calculated, since IRT true scores will be on a similar scale as the raw scores, which relates IRT ability to true scores. In addition, everyone at the same ability level θ has the same number-right true score. For dichotomous items, Lord and Novick (1968) defined a person's number-right true score ξ as the expectation of his observed score x, which is calculated as

$$\xi = \sum_{i=1}^{n} P_i(\theta) \tag{5}$$

For polytomous items, this formula can be extended as

$$\xi = \sum_{i=1}^{n}\sum_{k=1}^{m_i} W_{ik}\, P_{ik}(\theta) \tag{6}$$

where *n* is the total number of items on a test; item *i* has m_i response categories 1, 2, … m_i; W_{ik} is the weight (or score) associated with the k^{th} response category of item *i*; and $P_i(\theta)$ and $P_{ik}(\theta)$ are the item response function and item category response function for item *i* for a dichotomous IRT model and a polytomous IRT model, respectively.

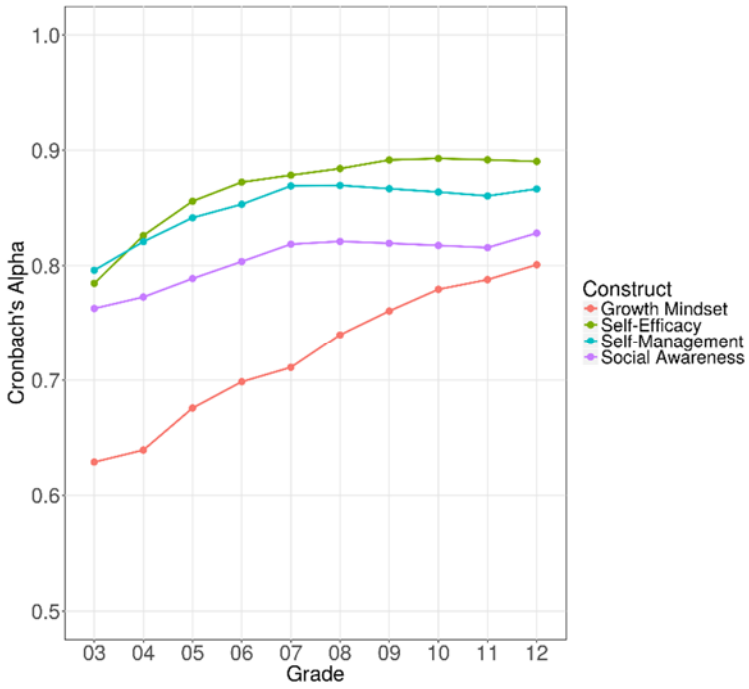
Results

Although the focus of this paper is on IRT modeling, we first present results from classical item analysis, factor analysis, and DIF analysis, followed by IRT analysis results.

Classical Item Analysis

For each SEL construct and at each grade level, we calculated the reliability as measured by Cronbach's alpha. These reliability coefficients are shown in Figure 2.

Figure 2. Cronbach's Alpha Coefficients of the SEL Constructs at Each Grade Level



The reliability of the self-management, self-efficacy, and social awareness scales are relatively high, ranging between 0.76 and 0.89. Generally speaking, the reliability coefficients are higher at higher grades. However, the reliability coefficients of the growth mindset scale, especially below grade 7, are lower than 0.7. As 0.7 and above is the recommended level of reliability estimates for student survey measures, this indicates less ideal internal consistency of the growth mindset scale when administered to students below grade 7. It is worth noting that our recent research has demonstrated the potential of increasing the reliability of SEL scores—

especially for lower grade levels—via subscore augmentation techniques, which allow for incorporating collateral information from the entire SEL survey (see Wang, Meyer, & Rice, 2018 for more details).

Under the CTT framework, we also conducted analyses at the item level. Table 4 shows descriptive statistics for each SEL item administered to 8th graders in 2015-16.

Table 4. Descriptive Statistics of SEL Items (Grade 8)

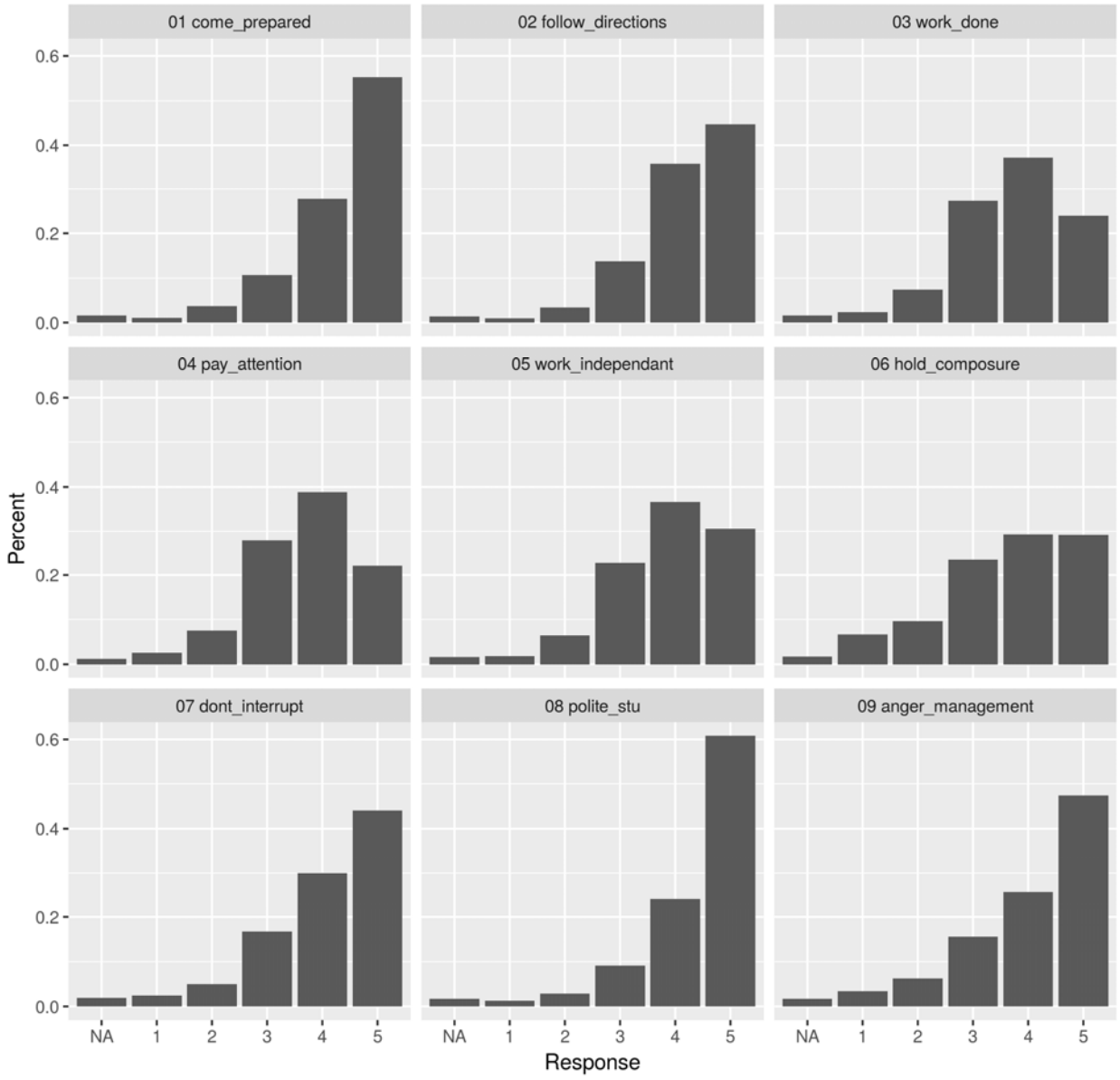
Item	Mean	Std. Dev.	Skewness	Kurtosis
Self-Management				
1	3.34	0.89	-1.41	1.61
2	3.22	0.88	-1.09	0.96
3	2.74	0.99	-0.54	-0.09
4	2.71	0.98	-0.55	-0.01
5	2.89	0.98	-0.67	-0.02
6	2.66	1.19	-0.63	-0.45
7	3.10	1.02	-1.06	0.58
8	3.43	0.87	-1.68	2.64
9	3.09	1.09	-1.11	0.45
Growth Mindset				
10	2.31	1.33	-0.11	-1.19
11	2.87	1.31	-0.80	-0.62
12	2.72	1.24	-0.65	-0.62
13	3.14	1.16	-1.21	0.43
Self-Efficacy				
14	2.69	1.19	-0.60	-0.57
15	2.33	1.14	-0.24	-0.73
16	2.14	1.21	-0.11	-0.89
17	2.60	1.12	-0.42	-0.65
Social Awareness				
18	2.78	0.91	-0.80	0.71
19	2.88	1.12	-0.99	0.33
20	2.51	1.09	-0.56	-0.27
21	2.77	0.96	-0.78	0.43
22	2.20	1.16	-0.29	-0.66
23	2.60	0.99	-0.61	0.08
24	2.54	1.09	-0.56	-0.24
25	2.38	1.10	-0.39	-0.45

All items are on a 5-point scale, so the raw scores range from 0 to 4. Since growth mindset items are negatively phrased, those responses were reverse-coded so that high scores on all items are in the same direction. Table 4 shows that, on average, the variance of growth

mindset items is the highest among the four constructs, which indicates that some students could have misunderstood the negative wording and selected the opposite end of the item responses.

For all items, students tend to select the response which reflects high social-emotional skills. Table 4 shows that all items have a mean score that is above 2. Item 8 on the self-management scale has the highest mean score of 3.43 (It asks whether the student was polite to adults and peers during the past 30 days). Figure 3 shows the responses of item 8 along with other items on the self-management scale for students in grade 8. In responding to item 8, almost 90% of the students answered “Almost all the time” (5) or “Often” (4); very few students selected “Almost never” (1), “Once in a while” (2), or “Sometimes” (3). Overall, almost all items on the SEL survey are negatively skewed, yet they exhibit somewhat reasonable spread, which allows the items to distinguish among students with high and low social-emotional skills.

Figure 3. Histograms of Items on the Self-Management Scale, Grade 8



We further examined each item's difficulty (i.e., *p*-value) and discrimination (i.e., polyserial correlation) indices. Overall, the SEL items are relatively easy, but the item discrimination coefficients are good, which shows that items within each construct seem to measure the same construct.

The average *p*-values and polyserial correlation coefficients across all items within each SEL construct and across all grades are summarized in Table 5. On average, self-management items are the easiest to endorse, and self-efficacy items are the hardest to endorse. Also, self-efficacy items can best differentiate students with high and low skills within the measured construct among the four SEL measures. Overall, all items have relatively high item discrimination indices (i.e., higher than 0.60).

Table 5. Average *p*-values and Polyserial Correlation Coefficients across All Items and Grades

	Item Difficulty (*p*-value)	Item Discrimination (Polyserial Correlation)
Self-management	0.77	0.77
Growth mindset	0.69	0.82
Self-efficacy	0.63	0.90
Social awareness	0.68	0.71

As mentioned above, the research sample excluded students who skipped more than 50% of the items within each SEL measure. We describe missing responses in the full sample in this section. Percent of missing responses on a single item ranged from 1% to 10%. Averaging across all grades, growth mindset items had the highest missing rate (3.0%). Self-efficacy items had the lowest missing rate (2.0%). In addition, the missing pattern varied across grades and constructs. As shown in Table 6, percent missing in growth mindset items varied across grades the most among the four constructs. The average missing rate among 3rd graders was 6.7%; this missing rate dropped to 2.9% among 7th graders and 1.9% among 12th graders. Percent missing varied somewhat in self-management and self-efficacy items, but not as dramatically as in growth mindset. Missing in social awareness is relatively consistent across grades. This could result from (i) confusion regarding the meaning of negatively phrased growth mindset items among young students and (ii) from social awareness items being located at the very end of the SEL survey, suggesting students may have lost interest or run out of time towards the end of the survey.

Table 6. Average Percent Missing across Items, by SEL Construct and Grade Level

	Percent Missing			
Grade	Self-Management	Growth Mindset	Self-Efficacy	Social Awareness
3	4.7%	6.7%	3.8%	4.5%
4	3.3%	4.6%	2.6%	3.7%
5	2.3%	3.3%	2.0%	2.8%
6	2.2%	3.2%	1.8%	4.4%
7	2.2%	2.9%	2.1%	4.7%
8	1.9%	2.4%	1.6%	4.3%
9	1.8%	2.3%	1.7%	4.3%
10	1.5%	2.1%	1.5%	4.0%
11	1.4%	1.9%	1.3%	3.5%
12	1.5%	1.9%	1.4%	3.6%

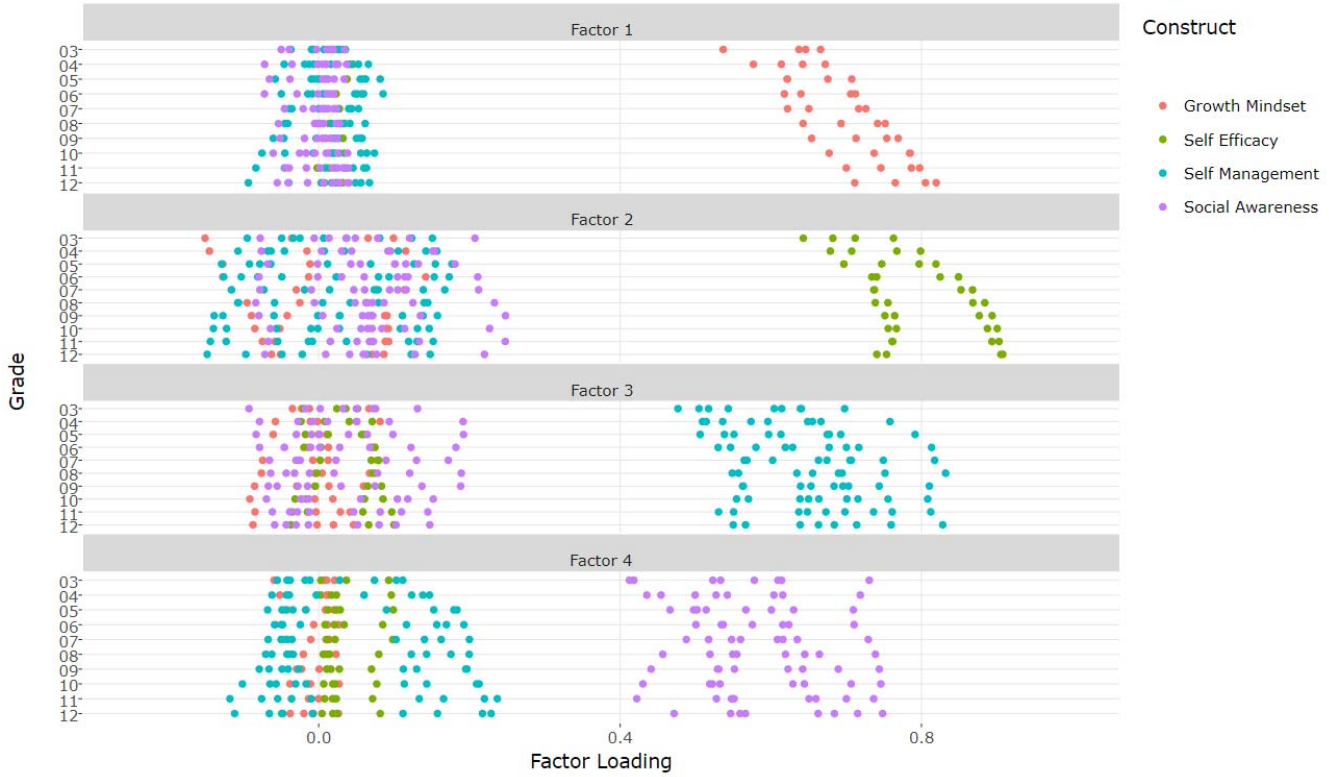
More importantly, percent missing was much higher when looking across items within a construct. Under the CTT framework, these missing items have to be excluded in calculating sum or mean scores for the corresponding construct. For example, students could have the

same mean score on an SEL construct, despite having answered different items (hard or easy) and different numbers of items. One advantage of IRT is to take into consideration the items a student skipped and use pattern scoring to create scale scores, which entails more precision. This will be discussed further in the IRT modeling section.

Factor Analysis

Figure 4 shows the factor loadings of a four-factor exploratory factor analysis at each grade level from grade 3 to grade 12. It is clear that items within each construct had high loadings on its own factor, as intended. No item was found to cross load onto another factor. These results support the use of unidimensional IRT models for each construct, which is discussed in a later section.

Figure 4. Factor Analysis Results



DIF Analysis

DIF analysis enables us to investigate whether students from different demographic subgroups vary in how they responded to the survey questions, especially when the subgroups have similar overall social-emotional skills at the measured domain. We conducted a total of 160 DIF runs between four pairs of demographic subgroups—male vs. female, White vs. African American, White vs. Hispanic/Latino, and English language learns vs. their peers—for each

construct and at each grade level (note that we are also in the process of conducting additional DIF analyses between grades for IRT scaling). Using the DIF flagging criteria presented above, we summarized the items exhibit category C DIF in Table 7.

Table 7. Item Numbers and Grades (in parenthesis) Exhibit Moderate to Large DIF (Category C)

Construct	Comparison Groups	Item (Grade)
Self-Management	Male vs. Female	-
	White vs. African American	3 (Grades 8, 11)
	White vs. Hispanic/Latino	-
	ELL vs. Others	3 (Grades 11, 12)
Growth Mindset	Male vs. Female	-
	White vs. African American	-
	White vs. Hispanic/Latino	-
	ELL vs. Others	-
Self-Efficacy	Male vs. Female	-
	White vs. African American	-
	White vs. Hispanic/Latino	-
	ELL vs. Others	-
Social Awareness	Male vs. Female	-
	White vs. African American	19 (Grade 10)
	White vs. Hispanic/Latino	-
	ELL vs. Others	-

Only two items (item 3 and item 19) were found to have moderate-to-large DIF across all DIF analyses conducted. Item 3 asks students to respond to the following prompt: "During the past 30 days, I got my work done right away instead of waiting until the last minute." This item disfavors White students in the White vs. African American DIF analysis at grades 8 and 11. It also disfavors non-ELL students at grades 11 and 12. Item 19 asks students, "During the past 30 days, how much did you care about other people's feeling?" It disfavors African American students in the White vs. African American DIF analysis at grade 10. An item exhibiting DIF does not necessarily mean that it is biased. However, these items should be carefully reviewed by SEL content experts before the next survey administration.

IRT Modeling

A main goal of this paper is to apply IRT modeling approaches to understand the measurement properties of CORE's SEL items. As discussed in the Analyses section, we used three polytomous IRT models (i.e., PCM, GPCM, NRM) for each of the four SEL constructs and each of the 10 grade levels. With a total of 120 separate IRT item calibrations, we highlight results that are most related to the questions that motivated this study.

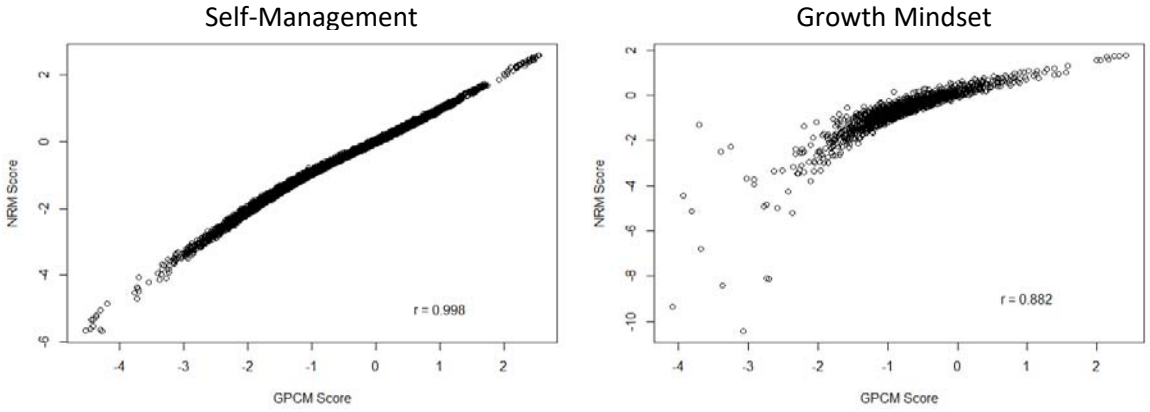
Hambleton, Swaminathan, and Rogers (1991) stated that "the advantages of item response models can be obtained only when the fit between the model and the test data of interest is satisfactory" (p. 53). When the model fits the data, the IRT model is useful in that it can accurately reflect the data. A poorly fitting IRT model will not result in accurate estimates of item and ability parameters. After conducting separate IRT item calibrations for each grade and each SEL construct, we first compared the model-data fit of the three polytomous IRT models—PCM, GPCM, and NRM. Generally speaking, the more general the model is, the better the model fit is. In other words, NRM should have the best fit with the data, since it is the most general model among the three, allowing each response category to have its own slope and intercept/location parameters. GPCM should fit better than PCM but not as well as NRM, since it is less general by providing each item's response categories with a constant slope while different items may have different slopes. PCM should fit the data the least well among the three, since the slope parameter is not estimated and every item has a fixed slope parameter of 1. We found the predicted pattern using the SEL data. Table 8 summarizes model-fit statistics of self-management at grade 7. It shows that, based on all model-fit statistics, such as AIC, BIC, log-likelihood, and chi-square, GPCM fits significantly better than PCM, and NRM fits significantly better than GPCM. The goodness-of-fit improved much more from PCM to GPCM than from GPCM to NRM.

Table 8. Model Fit Results of Self-Management Calibration (Grade 7)

	AIC	BIC	Log-Likelihood	χ^2	df	*p*
PCM	833598	833918	-416762			
GPCM	830093	830483	-415002	3521	8	<.001
NRM	828853	829478	-414355	1294	27	<.001

Correlations between IRT scale scores from the GPCM and NRM revealed that model selection could make a difference in students' scale scores. The figure below shows the correlation of GPCM and NRM scale scores for self-management and growth mindset scales at grade 7. The correlations for self-management, self-efficacy, and social awareness were high (i.e., above 0.99). However, the correlation coefficient for the growth mindset scale scores was 0.88; in other words, students who were low on growth mindset skills would be affected the most when a different IRT model is selected.

Figure 5. Correlation of GPCM and NRM Scale Scores for Self-Management and Growth Mindset Scales (Grade 7)



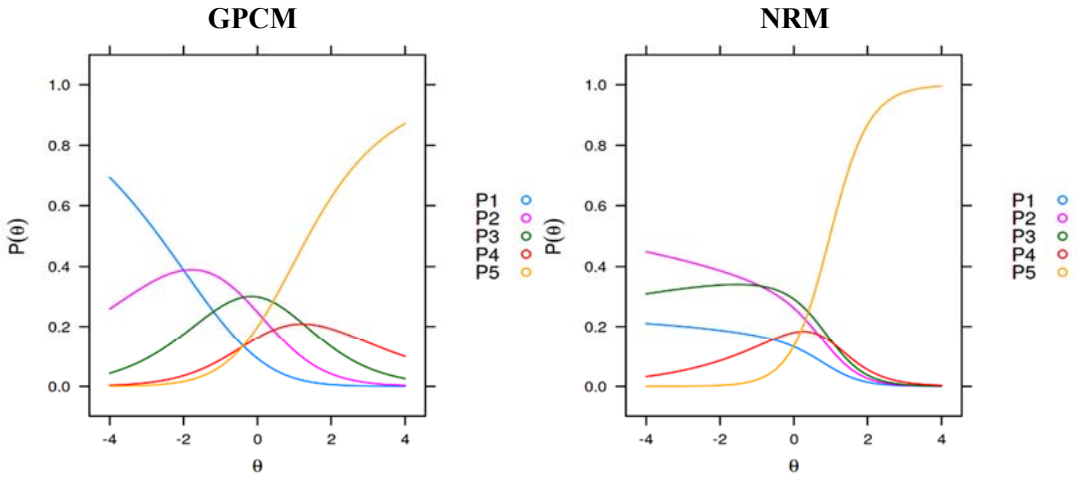
Although NRM fits the data statistically significantly better than the other two IRT models, this does not mean NRM should be used under all situations. Several factors need to be taken into account before making decisions on model selection. Besides a model's goodness-of-fit and item properties, the simplicity or parsimony of the model should also be considered. To achieve a balance between goodness-of-fit and model simplicity, PCM and GPCM, which are easier to work with than NRM, seems to be a better option for IRT calibration and scoring. The most general model, NRM, is valuable for research when investigating item functionalities and improving the SEL instruments.

In terms of using either PCM or GPCM for IRT calibration and scoring, strong guidance does not exist, nor is there a consensus on which model to choose. On the one hand, research often draws the conclusion that "there is little support for preferring the GPCM over the PCM as an analysis model" (Glas & Jehangir, 2014, p.112). On the other hand, there are reasonable arguments to choose either one of the IRT models. One argument for using GPCM over PCM regards the weighting GPCM provides based on the different discriminating power that items have (OECD, 2017); this means the model has better model-data fit and provides a more precise estimation of student ability. In contrast, PCM has attractive properties, including its simplicity, and the raw score is a sufficient statistic of the Rasch measure (Wright & Stone, 1979). From a survey design perspective, each item on the scale may present a different and unique aspect of the assessed construct, and thus require them to be weighted equally in calculating the total construct score. One approach that can help in the model selection process is examining the association between external data and different sets of IRT scale scores, which is one of the future studies we will be working on.

The advantage of NRM is especially apparent when examining an individual item's properties. Figure 6 is a comparison of the item category response function (ICRF) curves under GPCM and NRM for the first item on the growth mindset scale. This item has a relatively low slope parameter under GPCM, as shown in the flat curves on the left. Aside from that, the item displays adequate properties based on the GPCM ICRF plot. However, the NRM ICRF plot

reveals more problems with this item—three of the five slope parameters are negative, and one is close to zero. Students cannot distinguish between the response option of "Not At All True" and the other four response options (i.e., "A Little True," "Somewhat True," "Mostly True," and "Completely True").

Figure 6. Item Category Response Function Curves Modeled under GPCM and NRM for Growth Mindset Item 10



Recall that all growth mindset items are negatively phrased. Using negatively phrased items may increase the variance in survey responses, but survey researchers have shown that negatively phrased survey items can be difficult to understand for young children, especially when they are at the stage of learning classification and temporal relations. At this young stage, logical forms such as negations are still challenging to interpret (Benson & Hocevar, 1985; Gehlbach, 2015; Marsh, 1986). An examination of the interscale correlations of the observed raw scores at different grade levels also revealed much lower correlations between the growth mindset scale and the other three scales in lower grades compared to those in higher grades (see Table 9).

Therefore, the problems identified in this item could be due to its wording; students may simply have been confused and could not distinguish the nuances between the response options. Being able to identify this problem using NRM can help survey developers further modify and improve the measurement scale. Moreover, the use of IRT mixture models has demonstrated potential for identifying the degree to which some students may be confused in responding to questions with negatively worded items (see Bolt, Wang, Meyer, & Rice, 2018 for more details).

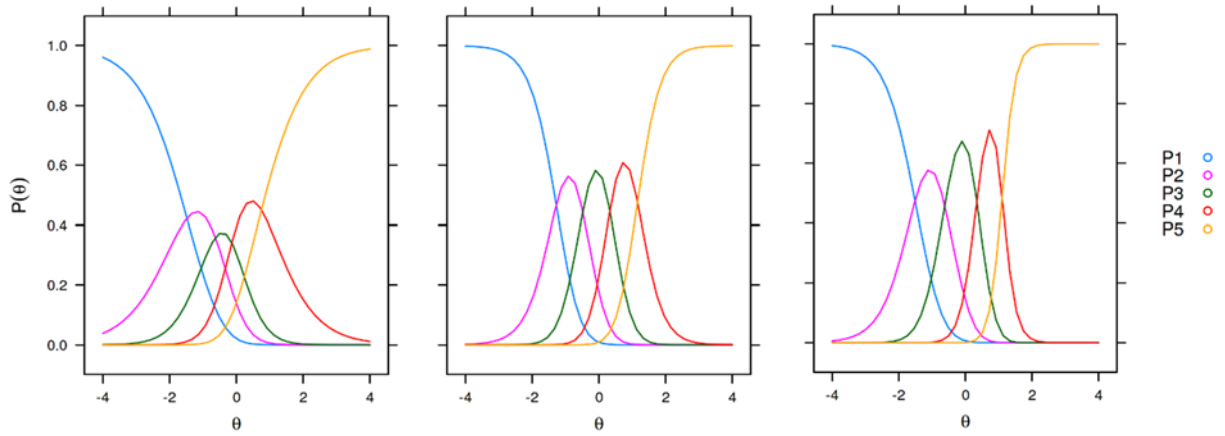
A close examination of each item's ICRFs shows that most items on the SEL instruments function properly. However, growth mindset items do not function well, especially among young students. This could be related to the negative phrasing. Relatedly, CORE has recently piloted positively phrased versions of the growth mindset items to consider using in the next SEL administration. More research on this issue should be expected in the future.

Table 9. Interscale Correlations between the Four SEL Scales in Grades 3, 7, and 11

Grade 3	Self-Management	Growth Mindset	Self-Efficacy	Social Awareness
Self-Management	1			
Growth Mindset	0.11	1		
Self-Efficacy	0.51	0.10	1	
Social Awareness	0.54	0.05	0.50	1
Grade 7	Self-Management	Growth Mindset	Self-Efficacy	Social Awareness
Self-Management	1			
Growth Mindset	0.23	1		
Self-Efficacy	0.49	0.30	1	
Social Awareness	0.60	0.16	0.49	1
Grade 11	Self-Management	Growth Mindset	Self-Efficacy	Social Awareness
Self-Management	1			
Growth Mindset	0.21	1		
Self-Efficacy	0.37	0.29	1	
Social Awareness	0.50	0.17	0.35	1

In addition, the ICRFs seem to support the hypothesis that students from different grades perceive items with the same wording differently. Figure 7 displays the ICRFs from the NRM for the same self-efficacy item administered to students in grades 4, 8, and 12. The item asks students how confident they are about the statement, "I can master the hardest topics in my classes." The response options are, "Not at all confident," "A little confident," "Somewhat confident," "Mostly confident," and "Completely confident." Figure 7 shows that students in grade 4 were less likely to select the middle response option, "Somewhat Confident;" fourth graders with average self-efficacy skills tend to select response the fourth option, "Mostly confident." In comparison, students at grades 8 and 12 with average self-efficacy skills tend to select the middle response option, "Somewhat confident." In addition, this item is more discriminating for 12th graders than for 8th graders. Students from lower grades and students from higher grades probably have different perceptions of "the hardest topics" in their classes, so it is not surprising that they respond to the same item differently. This finding also provides evidence that separate calibrations should be conducted at each grade level. In practice, although items with the exact same wording were administered to all students, item or scale scores derived from these items are not comparable across grades, because there is evidence that students from different grades perceive the same items differently.

Figure 7. An Example Self-Efficacy Item's ICRFs from the NRM at Grades 4, 8, and 12



Grade 4 Grade 7 Grade 12

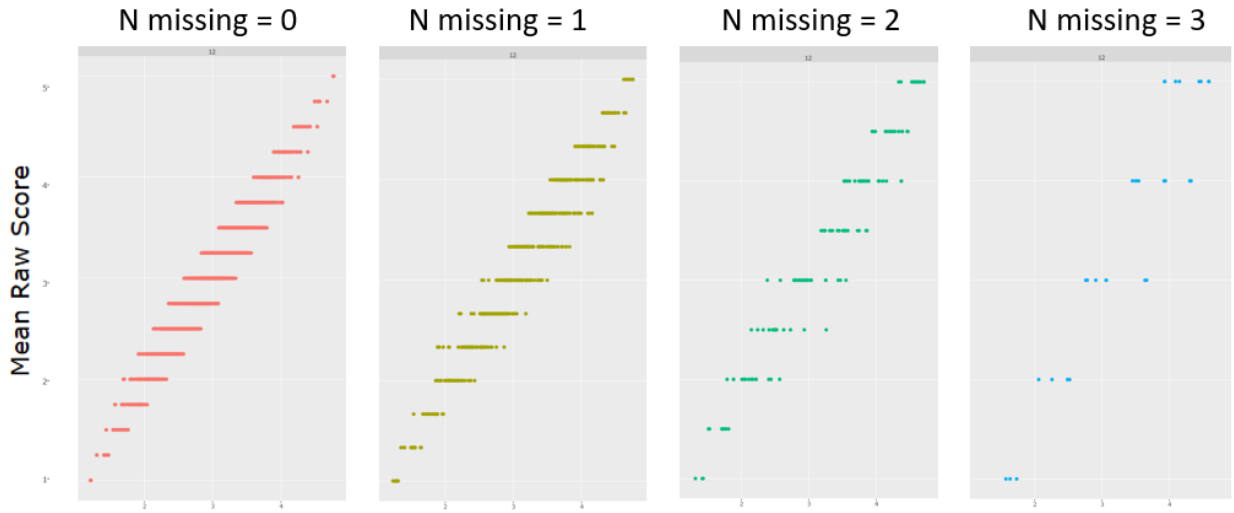
We also compared student outcomes estimated using polytomous IRT models to the classical approach (i.e., raw mean scores excluding missing). Missing responses have to be dropped under the classical approach. Consider a student who answered two of the four self-efficacy items. This student's raw score on self-efficacy is calculated as an average of the two items he or she answered, whereas most students' raw scores are calculated based on four items. In comparison to CTT, IRT has the advantage of handling missing responses by making use of the available response patterns to estimate a student's ability, while providing a larger standard error of estimate.

Figure 8 illustrates how missing responses affect score estimates under IRT PCM and GPCM true scores, and CTT raw scores using grade 12 growth mindset data. Figure 8a shows the relationship between PCM true scores and the raw scores. Because PCM is a type of Rasch model, estimated PCM scores and the observed raw scores have a one-to-one relationship when there is no missing. As the amount of missing data increases, the correlation between IRT true scores and CTT raw scores decreases. For example, when a student skips one item on the growth mindset scale, a raw score of 3 does not necessarily have an IRT true score of 3; depending on which item the student skipped and the difficulty of that item, the student may have an IRT true score that is different but relatively close to 3. When a student skips two or three items, a raw score of 3 could deviate further away from 3, depending on which items that student skipped.

Similarly, Figure 8b shows the relationship between GPCM true scores and the raw scores. Because GPCM weighs each item differently by its discrimination parameter, the same mean raw score can be associated with different GPCM true scores depending on the different combinations of items and responses. As the amount of missing data increases, the correlation between GPCM true scores and CTT raw scores decreases as well.

Figure 8. Plots of Mean IRT PCM/GPCM True Scores and CTT Raw Scores, by Number of Missing (Growth Mindset, Grade 12)

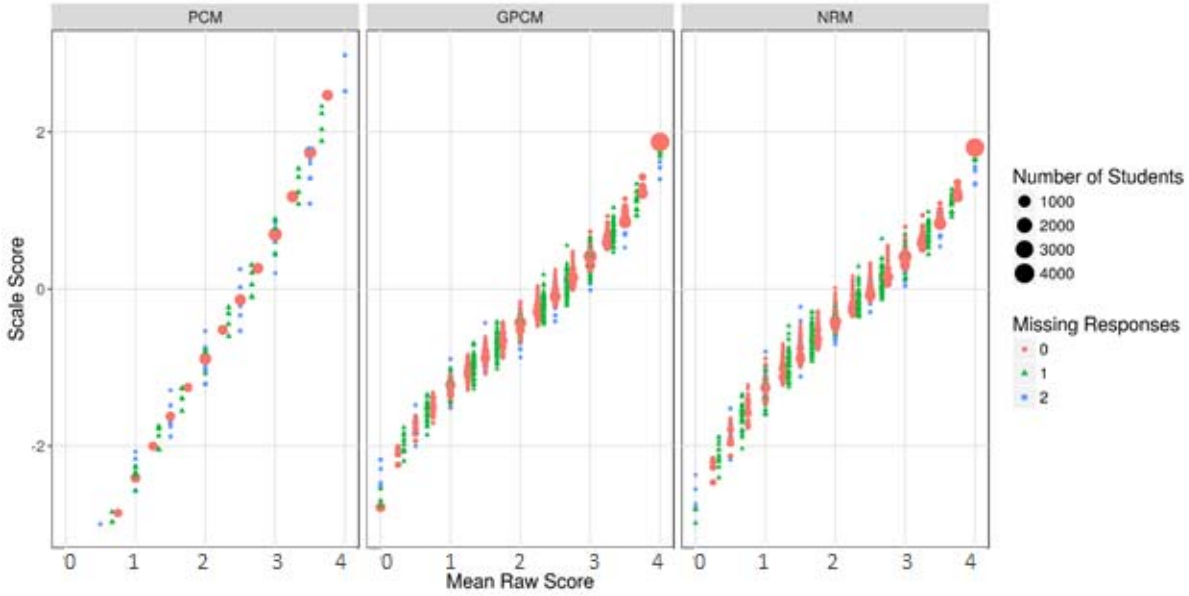
a. PCM Mean True Score



b. GPCM Mean True Score

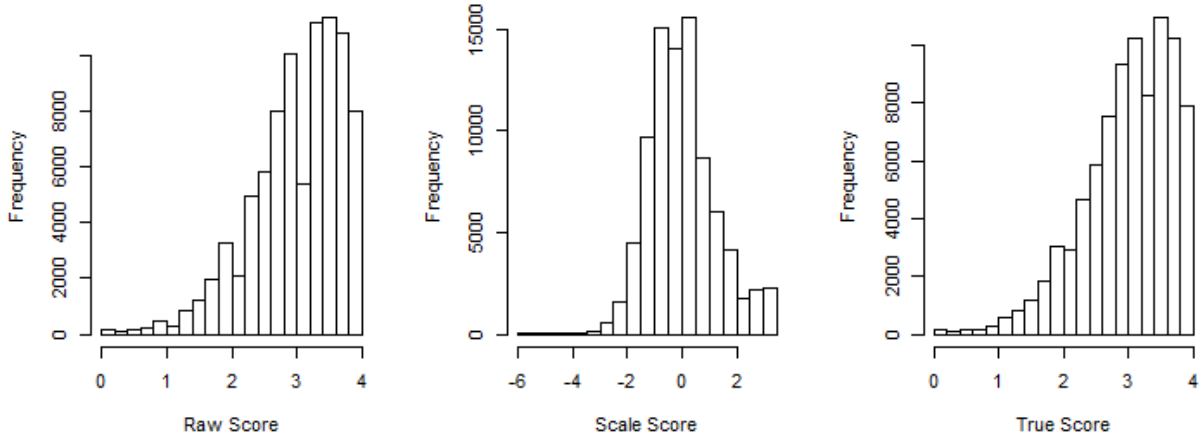
Figure 9 also shows how missing responses affect score estimates under IRT and CTT; it compares PCM, GPCM, and NRM scale scores to CTT raw scores side-by-side and highlights the number of students at each score point. It is difficult to handle missing responses in CTT, because missing responses cannot be properly scored unless they are dropped or imputed in some way. In comparison, when missing data are present, IRT can still perform calibration and scoring by using all of the available information based on the likelihood algorithm. Therefore, we suggest applying IRT models to properly score and understand the SEL data when missing data are present.

Figure 9. Plots of Scale Scores from PCM, GPCM, and NRM and CTT Raw Scores, by Number of Missing (Self-Efficacy, Grade 7)



In addition, we compared the distributions of the raw scores, IRT scale scores, and IRT true scores of each SEL construct to identify potential floor and ceiling effects. Such effects are often a concern raised in self-report survey responses, in which students simply select the responses associated with the highest or lowest ability levels/skills on all items, thereby resulting in a highest obtainable score or a lowest obtainable score. As expected, we found negative skewness in raw observed scores and true scores, yet, as shown in Figure 10, the use of IRT scale scores (middle panel) can help mitigate the skewness in raw score (left panel) and true score (right panel) distributions.

Figure 10. Distributions of Raw Scores, Scale Scores, and True Scores of Self-Management (Grade 8)



Conclusion

It is widely acknowledged that social-emotional skills play an important role in one's academic development, workforce performance, and well-being. However, social-emotional skills are still the missing piece in K-12 education and have not been extensively studied. As the first large-scale implementation in the country to assess students' SEL, the CORE Districts provide an unprecedented opportunity to study SEL measurement properties.

Results from classical item analyses show that the reliability of self-management, self-efficacy, and social awareness scales are relatively high. However, the reliability of growth mindset, especially at lower grades, is less ideal. Although further research is needed in the area, this could be attributed to negatively phrased items creating confusion among young students. In practice, growth mindset scores for students at lower grades should be interpreted with caution. In addition, DIF analyses detected a few items that have moderate-to-large DIF. Those items require closer examination before the next survey administration.

Results from factor analyses revealed a clear four-factor solution in which items are clustered as intended. Thus, different SEL constructs were calibrated separately in the subsequent IRT analyses. IRT modeling provides several advantages over the classical approach, including handling missing responses, recognizing differences in students' understanding across grades, providing proper weights in scoring by considering the difficulty and discrimination properties of the survey items, and providing comparable scale scores with test/survey content changes over the years. Therefore, we focused on using an IRT approach when examining the measurement properties of CORE's SEL instruments.

A comparison of three polytomous IRT models shows that the NRM fits the data statistically significantly better than the GPCM, which in turn fits the data better than the PCM. We also found that model selection makes a difference in students' ability estimates and item parameter estimates. To achieve a balance between goodness-of-fit and model simplicity, GPCM, which is easier to work with than NRM, is recommended for IRT calibration and scoring for CORE's SEL survey. The NRM is valuable for research to investigate item functionalities and to further improve the SEL measurements.

At the item level, results show that growth mindset items do not function well, especially among young students. Again, this could be related to the negative phrasing of these items. In response to these findings, and as part of their efforts towards continuous improvement, CORE has recently piloted positively phrased versions of the growth mindset items to consider using in the next SEL administration.

An examination of individual items' ICRFs provides evidence that students from different grades perceive items with the same wording differently. Therefore, even an item with the same wording should be treated as different items when administered to students at different grade levels. When we interpret the observed raw scores and IRT scale scores (before proper scaling is done), the scores from different grade levels are placed on different scales and

are not directly comparable. In future work, we plan to develop vertically scaled scores that will enable comparison of students' SEL scores across grades.

In terms of scoring CORE's SEL survey, we also showed that IRT models have an advantage over CTT by being able to handle missing responses and use pattern scoring to create scale scores, which is important to consider when missing data are present. IRT scale scores can also help alleviate some degree of the skewness in raw score and true score distributions.

In sum, the results presented in this paper show that, for the most part, CORE's SEL instruments have reasonable measurement properties. However, they have room for improvement. Specifically, negatively phrased items should be reworded; items exhibiting DIF require careful examination; current SEL items should be replaced with new items after a few years of administration; administering different sets of items to different grades should also be considered and explored. At this point, practitioners should be careful before interpreting and making inferences based on the survey results.

References

Allen, N. L., Donoghue, J. R., & Schoeps, T. L. (2001). *The NAEP 1998 Technical Report* (NCES 2001–509). Washington, DC: National Center for Education Statistics.

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

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

Battistich, V., Schaps, E., & Wilson, N. (2004). Effects of an elementary school intervention on students' "connectedness" to school and social adjustment during middle school. *Journal of Primary Prevention, 24*(3), 243–262.

Belfield, C., Bowden, B., Klapp, A., Levin, H., Shand, R., & Zander, S. (2015, February). *The economic value of social and emotional learning* (Revised Version). New York: Center for Benefit-Cost Studies in Education: Teachers College, Columbia University.

Benson, J. & Hocevar, D. (1985). The impact of item phrasing on the validity of attitude scales for elementary school children. *Journal of Educational Measurement, 22*, 231-240.

Bock, R. D. (1972). Estimating item parameters and latent ability when responses are scored in two or more nominal categories. *Psychometrika, 37*(1), 29-51.

Bolt, D. M., Wang, Y. C., Meyer, R. H., & Rice, A. B. (2018). I*RT 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.

Bridgeland, J., Bruce, M., & Hariharan, A. (2013). *The missing piece: A national survey on how social and emotional learning can empower children and transform schools*. Washington, DC: Civic Enterprises.

Chetty, R., Friedman, J. N., Hilger, N., Saez, E., Schanzenbach, D. W., & Yagan, D. (2011). How does your kindergarten classroom affect your earnings? Evidence from Project Star. *Quarterly Journal of Economics, 126*(4), 1593-1660. doi:10.1093/qje/qjr041.

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.

Cunningham, W., & Villaseñor, P. (2016). *Employer voices, employer demands, and implications for public skills: Development policy connecting the labor and education sectors*. Washington, DC: World Bank Group.

de Ridder, D. T., Lensvelt-Mulders, G., Finkenauer, C., Stok, F. M., & Baumeister, R. F. (2012). Taking stock of self-control: A meta-analysis of how trait self-control relates to a wide range of behaviors. *Personality and Social Psychology Review, 16*(1), 76–99.

Dorans, N. J. & Kullick, E. (1986). Demonstrating the utility of the standardization approach to assessing unexpected differential item performance on the Scholastic Aptitude Test. *Journal of Educational Measurement, 23*, 355-368.

Dorans, N. J., & Schmitt, A. P. (1991). *Constructed response and differential item functioning: A pragmatic approach* (Research Report No. RR-91-47). Princeton, NJ: Educational Testing Service.

Durlak, J. A., Weissberg, R. P., Dymnicki, A. B., Taylor, R. D., & Schellinger, K. B. (2011). The impact of enhancing students' social and emotional learning: A meta-analysis of school-based universal interventions. *Child Development, 82*(1), 405-432.

Dusenbury, L., Newman, J. Z., Weissberg, R. P., Goren, P., Domitrovich, C. E., & Mart, A. K. (2015). Developing a blueprint for preschool to high school education in social and emotional learning: The case for state learning standards. In J. A. Durlak, C. E. Domitrovich, R. P. Weissberg, & T. P. Gullotta (Eds.), *Handbook of social and emotional learning: Research and practice* (pp. 532-548). New York: Guilford Press.

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

Elementary and Secondary Education Act of 1965, 20 U.S.C. § 1111 (2015).

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

Glas, C. A. W., & Jehangir, K. (2014). Modeling country-specific differential item functioning. In L. Rutkowski, M. von Davier, & D. Rutkowski (Eds.), *Handbook of International Large-Scale Assessment: Background, Technical Issues, and Methods of Data Analysis*. (pp. 97-115). New York: Springer.

Hambleton, R. K., Swaminathan, H., & Rogers, H. J. (1991). *Fundamentals of item response theory*. Newbury Park, CA: Sage.

Hawkins, J. D., Kosterman, R., Catalano, R. F., Hill, K. G., & Abbott, R. D. (2008). Effects of social development intervention in childhood 15 years later. *Archives of Pediatrics and Adolescent Medicine, 162*, 1133–1141.

Heckman, J. J. (2008). Schools, skills, and synapses. *Economic Inquiry, 46*(3), 289–324. doi:10.1111/j.14657295.2008.00163.x

Heckman, J. J., Humphries, J. E., & Kautz, T. (Eds.) (2014). *The Myth of Achievement Tests: The GED and the Role of Character in American Life*. Chicago: University of Chicago Press.

Jones D. E., Greenberg M., Crowley M. (2015). Early social-emotional functioning and public health: The relationship between kindergarten social competence and future wellness. *American Journal of Public Health, 105*(11), 2283–2290.

Kolen, M. J., & Brennan, R. L. (2014). *Test equating, linking, and scaling: Methods and practices* (3rd ed.). New York: Springer.

Lord, F. M. & Novick, M. R. (1968). *Statistical Theories of Mental Test Scores*. Addison-Wesley, Menlo Park.

Mantel, N. (1963). Chi-square tests with one degree of freedom: Extensions of the Mantel-Haenszel procedure. *Journal of the American Statistical Association, 58*, 690-700.

Mantel, N., & Haenszel, W. (1959). Statistical aspects of the analysis of data from retrospective studies of disease. *Journal of the National Cancer Institute, 22*, 719-748.

Marsh, H. W. (1986). Negative item bias in rating scales for preadolescent children. A cognitive developmental phenomenon. *Developmental Psychology, 22*, 37-49.

Martin, M. O., Mullis, I. V. S., Hooper, M., Yin, L., Foy, P., & Palazzo, L. (2016). Creating and Interpreting the TIMSS 2015 Context Questionnaire Scales. In M. O. Martin, I. V. S. Mullis, & M. Hooper (Eds.), *Methods and Procedures in TIMSS 2015* (pp. 15.1-15.312). Retrieved from http://timss.bc.edu/publications/timss/2015-methods/chapter-15.html

Masters, G. N. (1982). A Rasch model for partial credit scoring. *Psychometrika, 47*, 149-174. doi:10.1007/BF02296272.

Meyer, R. H., Gawade, N., & Wang, Y. (2016). *Implications of differential item quality for test scores and value-added estimates*. Paper presented at the NCME annual conference, Washington, DC.

Moffitt, T. E., Arseneault, L., Belsky, D., Dickson, N., Hancox, R. J., Harrington, H., … Caspi, A. (2011). A gradient of childhood self-control predicts health, wealth, and public safety. *Proceedings of the National Academy of Sciences, 108*(7), 2693–2698. doi:10.1073/pnas.1010076108.

Muraki, E. (1992). A generalized partial credit model: application of an EM algorithm. *Applied Psychological Measurement, 16*(2), 159-176.

Oakland Unified School District. (2015). *Quality School Development School Performance Framework Guidebook, 2015-2016 Edition*. Oakland, CA: Office of Post-Secondary Readiness, Oakland Unified School District. Available at: http://schoolperformanceframework.weebly.com/uploads/7/3/4/2/73421749/spf_guidebook_v7.5.2.pdf

Organisation for Economic Co-operation and Development. (2017). *PISA 2015 technical report*. France: Author. Available at http://www.oecd.org/pisa/data/2015-technical-report/.

Pearson. (January 10, 2017). *PARCC: Final technical report for 2016 administration*. New York: Author.

Poropat, A. E. (2009). A meta-analysis of the five-factor model of personality and academic performance. *Psychological Bulletin, 135*(2), 322-338.

Preston, K., Reise, S., Cai, L., & Hays, R. (2011). Using the nominal response model to evaluate response category discrimination in the PROMIS emotional distress item pools. *Educational and Psychological Measurement, 71*(3), 523-550.

Schulz, W., Ainley, J., & Fraillon, J. (Eds.) (2011). *ICCS 2009 Technical Report*. Amsterdam: IEA.

Segal, C. (2013). Misbehavior, education, and labor market outcomes. *Journal of the European Economic Association, 11*(4), 743-779. doi: 10.1111/jeea.12025.

Sharp, C., Steinberg, L., Yaroslavsky, I., Hofmeyr, A., Dellis, A., Kincaid, H., et al. (2012). An item response theory analysis of the Problem Gambling Severity Index. *Assessment, 19*, 167–175.

Smarter Balanced Assessment Consortium. (June 23, 2016). *Smarter Balanced Assessment Consortium: 2014-15 technical report*. Los Angeles: Author.

Taylor, R. D., Oberle, E., Durlak, J. A., & Weissberg, R. P. (2017). Promoting positive youth development through school-based social and emotional learning interventions: A meta-analysis of follow-up effects. *Child Development, 88*(4), 1156-1171. doi:10.1111/cdev.12864.

Wang, Y. C., Gawade, N., & Meyer, R. H. (2015). *Assessment properties and value-added measurement of educator effectiveness*. Paper presented at the NCME annual conference, Chicago, IL.

Wang, Y. C., Meyer, R. H., & Rice, A. B. (2018). *Incorporating collateral information for reporting scores of social-emotional learning measures*. Paper presented at the National Council on Measurement in Education annual conference, New York, NY.

West, M. R., Buckley, K., Krachman, S. A., & Bookman, N. (2018). Development and implementation of student social-emotional surveys in the CORE Districts. *Journal of Applied Developmental Psychology, 55*, 119-129.

Wright, B. D., & Stone, M. H. (1979). *Best Test Design*. Chicago, IL: MESA Press.