Published version of manuscript can be found here:

https://www.sciencedirect.com/science/article/pii/S019339731630137X

A Continuous Improvement Approach to Social and Emotional Competency Measurement

Laura A. Davidson[a]

Marisa K. Crowder[a,b]

Rachel A. Gordon[d]

Celene E. Domitrovich[c,e]

Randal D. Brown[a,b]

Benjamin I. Hayes[a]

[a]Washoe County School District

[b]University of Nevada, Reno

[c]CASEL: The Collaborative for Academic, Social, and Emotional Learning

[d]University of Illinois at Chicago

[e]Georgetown University

Author Note

This work was supported by the Institute of Education Sciences, U. S. Department of Education

We are deeply indebted to the full Researcher-Practitioner Team for their help with this project: Ann McKay-Bryson, Robert Schamberg, Trish Shaffer, Jeremy Taylor, and Roger Weissberg

Abstract

A four-year research project by a researcher-practitioner partnership to develop a self-report measure of student social and emotional competency to identify at-risk students and guide practice is described. A continuous measure improvement approach facilitated work towards this goal. Items developed collaboratively by the partnership team were administered annually to all 5th, 6th, 8th, and 11th graders in WCSD in the context of their online school climate survey. Despite strong construct validity, initial Rasch analysis revealed a substantial ceiling effect inhibiting assessment of students at the mid-to-high range of social and emotional ability. Student focus groups informed by a latent class analysis were conducted and expert practitioners and scholars refined the items. The process resulted in an improved measure as well as more consistent district-wide survey administration. Strengths and challenges of the scale development process and data use strategies are discussed along with recommendations for future assessment development efforts.

Keywords: researcher-practitioner partnership, social emotional competency, item response theory, continuous measure improvement, social emotional learning

A Continuous Improvement Approach to Social and Emotional Competency Measurement

In recent years, increased attention has been paid to the importance of students' social and emotional competencies (SECs), or the knowledge, skills, and attitudes needed to be personally and socially competent. The enhanced focus is the result of research documenting that students with SEC perform better in school, are more likely to stay in school and graduate, and function at higher levels in their adult lives than students without SEC (e.g., Farrington, Roderick, Allensworth, Nagaoka, Keyes, Johnson, & Beechum, 2012; Jones, Greenberg, & Crowley, 2015; Valiente, Swanson, & Eisenberg, 2012; see also Schamberg et al. in press). A lack of SEC also appears to be a risk factor for poor outcomes. Students who enter school with lower SECs have been shown to fall behind their peers in early elementary school and are at greater risk in adolescence for social adjustment problems, academic failure, and drop out (Arsenio, Adams, & Gold, 2009; Barry & Reschly, 2012; Domitrovich, Durlak, Staley, & Weissberg, 2017).

Given the importance of SEC, particularly for school success, it is surprising that relatively less attention has been devoted to assessment of these competencies. McKown (2016) provides several recommendations for ensuring that measures intended to assess SECs in educational contexts are high quality and useful. These include, among others, the need for SEC assessments to be: 1) conducted on a large scale without the need for trained clinicians or researchers, 2) based on strong theoretical models, 3) informed by educators so that they are practical and solve "real world" problems that teachers care about, and 4) able to assess a range of dimensions that can develop a comprehensive picture of a student's social and emotional needs and strengths.

In line with these recommendations, current Standards for Educational and Psychological Testing call for an approach to measure development that views reliability and validity as fluid properties that vary across populations, locations, and time, rather than fixed traits of instruments (American Educational Research Association, American Psychological Association, & National Council on Measurement in Education, 2014). The Standards also encourage test adopters to weigh reliability and validity evidence with the intended use of the measure in mind, and emphasize the importance of considering the applicability of existing evidence for local contexts, the accumulation of new evidence, where needed, and the periodic revision of tests. This kind of iterative test development and refinement process has been common for standardized tests that assess academic constructs, which are often developed by testing companies who regularly re-examine and refresh item pools and accumulate evidence through multiple strategies including large norming samples as well as cognitive interviews with test takers and content reviews by experts. Although testing companies do not always implement this continuous improvement approach completely and successfully, scrutiny from stakeholders and competition for large contracts has encouraged its intentional adoption throughout the test development, delivery, and revision process (Wild & Ramaswamy, 2008).

In the area of SECs, greater attention is spawning new assessments, and existing measures are sometimes developed with, or disseminated by, testing and publishing companies. Yet, historically, SEC measures have also frequently been developed by independent scholars or research teams, sometimes reflecting substantial content knowledge, but not the continuous measure improvement approach discussed above. Many developers of these measures have aimed for a fixed and relatively small list of items, with scoring based on simple item sums or averages, and revisions avoided to maintain cross-time comparability (see Denham, 2016 for a description of several established SEC measures). As a result, assessments are rarely adapted to meet the needs of local contexts. Past reliance on classical test theory approaches has also focused on test-level information for reliability and validity evidence (e.g., Cronbach’s alpha). In contrast, the IRT approaches more common in testing firms offer detailed item-level statistics, support large item banks, and readily link old and new item sets through common items (Gordon, 2014, 2015). Leveraging these strengths of IRT can facilitate continuous improvement, including adjusting to school contexts where districts regularly re-administer surveys to students and may want to tailor items to local needs (including evolving student demographics, applicability across grade levels and to state-, district- or school-standards and pedagogy, and desire to present students with fresh items rather than the same small set year after year).

The project described in this paper applied the continuous measure improvement approach to the development of a student self-report rating of social and emotional competence. Implementation of our approach relied on a researcher-practitioner partnership, which bridged the psychometric expertise of academic scholars with the on-the-ground research and practice expertise in the school district. This project used a four-year, mixed-method, iterative approach to identify and address the key measurement challenges of the instrument, focusing on the intended goal of the partnership: to assess a large population of students with a range of SEC ability so that early intervention with students who need instructional support can occur. Rasch-based analyses of student survey items provided item-level statistics, latent class analyses identified groups of students that contributed to an identified ceiling effect, and focus groups explored reasons for the ceiling effect and helped refine items and the survey environment. This approach also offered local validity evidence by developing and refining items in relation to school district standards for SEC instructional practice and with input from teachers and students. The project also emphasized a continuous improvement approach, using item-level statistics based on analyses of each year’s survey to iteratively improve the item bank, ensure that items covered the full range of each construct, and allow selection of item subsets to be used at particular grade levels and in repeated administrations.

The project resulted in the development of a bank of 138 student self-report items that assess eight subdimensions of SECs. From this larger bank of items, two subsets (which we refer to as instruments in order to facilitate communication with internal and external audiences) were also created. The 138 item bank allows for the flexible creation of subsets for various uses, offers extensive examples of indicators aligned to the five SEC clusters defined by the Collaborative for Academic, Social, and Emotional Learning (CASEL), and has been useful for training purposes and cross-walking to learning standards. The first instrument includes 40 items that the partnership selected because they covered all subdimensions of interest, were well-aligned with district standards for SEC instructional practice, and performed well in iterative administrations with Washoe County School District (WCSD) students. The second is a 17-item instrument that the partnership felt would be useful as a short-form assessment of global SEC for when districts needed fewer than 40 items due to concerns about time and burden. This short-form measure is currently in use by several districts in stand-alone form or as part of a larger climate survey (see Washoe County School District & Collaborative for Academic, Social, and Emotional Learning, 2016). The full item bank and both the 17- and 40-item instruments are free, open-source and available for other districts to use and adapt for their measurement needs. This paper focuses on two subdimensions from the larger item bank, Relationship Skills (RS) and Self-Management of Emotions (SME), to illustrate the collaborative process undertaken by the partnership to address the ceiling effect found in the original measure and to produce an assessment of student social and emotional competence with evidence of local reliability and validity.

The WCSD-CASEL Research-Practice Partnership Project

CASEL and the NoVo Foundation developed the Collaborating Districts Initiative (CDI) to build the capacity of eight large urban school districts, including WCSD, to systematically provide support for social and emotional learning (SEL) to all preK-12 students and the adults who served them. SEL is the process through which students acquire and apply the knowledge, skills, and attitudes needed to be personally and socially competent. The goals of the CDI are to demonstrate that a district can implement SEL with fidelity, at scale across a system, as well as to collaboratively develop and refine practical tools that promote the effective implementation and assessment of SEL, and strengthen the research base related to this work. All districts in the CDI participated in a national evaluation conducted by the American Institutes for Research (AIR) which included the collection of student self-report ratings of social and emotional competence (American Institutes of Research & Collaborating Districts Initiative, 2014). It was these ratings that were iteratively refined in the current project.

In 2009, WCSD had developed an Early Warning Risk Index (EWRI) to help identify and intervene with students at risk for not graduating based on factors that paralleled Balfanz's original academic risk prediction model (Balfanz, Herzog, & Mac Iver, 2007; Balfanz & Byrnes, 2006), including credit deficiencies, test scores, absenteeism, suspensions, transiency, and retention. WCSD conducted a longitudinal analysis of the EWRI's effectiveness in predicting graduation rates for a sample of students who were 9th graders in the district in 2009. Results showed that only one-third of all "High Risk" 9th grade students beat the odds and graduated four years later, compared with 79% of students identified as "No Risk" in the 9th grade. Although impressed at its predictive utility, WCSD developed an interest in the 34% of high risk students who had some resilient characteristics unidentified by the model that allowed them to overcome substantial academic obstacles in the 9th grade and graduate four years later.

As a result of participating in the CDI, WCSD had collected the self-report SEC ratings mentioned above and wanted to explore whether SEC served as a protective factor that contributed to student resilience. Ultimately, WCSD intended to include the measure as part of its interactive online data reports which include the EWRI scores alongside academic and behavioral data to allow for improved educational decision-making. CASEL intended to use the measure to disseminate nationally with guidance to districts interested in using the assessment as part of their SEL measurement systems. Given these intended uses, both organizations wanted to deepen and broaden the evidence base about the measure.

The district data on student SEC, collected through its participation in the CDI evaluation, showed the subdimensions of the measure had moderate reliabilities ($\alpha = .68$ to .74), but also substantial ceiling effects. For example, 18% of WCSD students rated themselves at the highest level on all four items (*Very True for Me*) in the Social Awareness subdimension, and 14% rated themselves at the highest level on all seven items of the Self-Management subdimension. High self-ratings made it statistically difficult to use this rating to differentiate students at the mid-to-high level of social and emotional ability. Because developing a risk and resiliency model that could accurately identify students in need of intervention and support was a primary goal of the district, it was of critical importance that the SEC items could measure student abilities across all levels of functioning. Statistically, the ceiling effect limited the instrument’s ability to measure students’ growth over time, particularly for students who already self-reported moderate to high SEC. The lack of items measuring moderate and highly competent students also impeded the ability to correlate student social and emotional competencies with outcomes by producing higher standard errors of measurement and lower reliability for students whose higher SEC abilities were not adequately captured by the items.

Improving the student self-report SEC items became the focus of this four-year partnership, and the primary measurement challenge undertaken was to reduce the ceiling effect. Ideally, there would have been more items that were difficult for a student to endorse or an expanded set of response categories on the high end (e.g., *Always true for me)* that would be harder to endorse. Further, the addition of more items that were easier to endorse, or an expanded set of response categories on the lower end was needed. Table 1 shows key item information for each year, including metrics about the ceiling and floor effect. Figure 1 presents the three phases of the project and the approach taken to address this statistical challenge, including the lessons learned at the conclusion of each phase.

CASEL and WCSD used an Institute of Education Sciences Research-Practice Partnership grant to build upon their CDI work and improve the CDI self-report assessment of student SEC so that it would meet the district’s more specific needs (Schamberg et al., in press). A more reliable and valid measure of SEC would also address existing gaps in the field of social and emotional measurement. This partnership was consistent with the definition put forth by the William T. Grant Foundation in that it was a long-term, equally beneficial collaboration between researchers and school district practitioners that was designed to produce a solution to a problem of practice (Coburn, Penuel, & Geil, 2013). As described in this paper and elsewhere, the members of the partnership team were extremely committed to “two-way street” research (Tseng, 2012), meaning they developed strategies to intentionally engage all members of the partnership equally in the development process (Schamberg et al., in press).

The continuous measure improvement process described in this paper greatly benefited from the researcher-practitioner partnership that provided the context for the work. The scale development relied equally on the expertise of both practitioners deeply immersed in on-the-ground SEL implementation and district-level research and data use, and researchers who had a broad knowledge of the larger field of SEC measurement and SEL best practices. Perhaps most importantly, students and educators were partners in the work at all stages, providing feedback on the instrument, the survey environment, and the data patterns the measurement tool helped produce. This partnership approach was a critical ingredient that helped reduce the substantial gulf that often exists between psychometricians who primarily develop survey instruments, and practitioners, who use data to guide decision-making. Because the partnership created a group with diverse expertise and a structured approach to sharing knowledge with one another, the project led not only to a more defensible and useful instrument, but also better psychometric tools, more thoughtful data-sharing approaches with educators, and better strategies for partnering with students in school improvement efforts.

SEC Survey Administration

WCSD administers an online student climate and safety survey in the spring of each academic year that is taken by all students in grades 5 through 9, and 11. Half of all students complete the Safety Survey, which asks about school safety, bullying, and risk-taking behaviors, and half complete the Climate Survey. The SEC items are embedded within the Student Climate Survey, which also asks students 40-50 additional questions about the climate of their school, including their attitudes towards their education, peers, and school staff. Table 1 reports the number of students who received the SEC items, the response rate on the items, and how many items were tested each year of the project.

WCSD was already in its third year of implementing this annual district-wide climate survey to students, staff, and parents when the project began. WCSD also regularly disseminated climate survey data to educators as part of a larger book of school data and as part of a broad system of data-sharing events each year (described in more detail later in this article). As such, the district was primed for the continuous measure improvement process because the SEC items could be embedded into its existing survey administration process, and built-in opportunities and events for educators to reflect on the data had already been established. The district's capacity to use Qualtrics (2015) survey software also facilitated complex randomizations of different sets of items to students, allowing for a large bank of items to be tested within a single administration.

WCSD's survey administration procedures were approved annually by the district's legal team and institutional review board. A passive parental consent process was used, and the consent form notified parents that their students' identification numbers would be attached to their child's responses. When the student self-report SEC items were added to the climate survey, the consent form explained that the purpose of using student identifiers was to study the relationship between social and emotional competencies and academic and behavioral outcomes. Once data were collected and achievement and demographic data were linked to student responses, ID numbers were stripped from the file so student responses could not be traced back to them. Students completed the Student Climate Survey between April and June each year via computer at their school site. All schools were provided a brief proctoring script to read before each administration to ensure that consistent directions were utilized across administrations. Formal data on the number of passive consent forms returned by parents opting their children out of the Climate Survey is not currently collected by WCSD. Schools are instructed to keep returned passive consent forms at their site to ensure they know which students should not take the survey on the scheduled administration day. We contacted a random sample of four schools to determine how many opt-outs they received for the 2016 administration. Responses ranged from zero at one elementary school to six at another elementary school. One high school responded “a handful each year, fewer than five”, while the middle school responded that only 3-4 were received in the previous year. Though a small sample, these match other anecdotes from prior years of administration indicating that opt-outs from parents are very rare.

Analytic Approach

Item Response Theory (IRT) with Rasch modeling was the primary analytic strategy used in this project to provide information about the psychometric properties of the student self-report instrument each year. IRT provided some advantages over Classical Test Theory, including its ability to provide extensive detail about not only test-level validity, but item-level validity and particularly the distribution of items across a latent dimension (like relationship skills). Classical Test Theory tends to focus primarily on the correlations among items within a latent dimension, but IRT encourages measurement developers to think about not only which items are relevant to a construct, but which level each item assesses within the dimension (Gordon, 2015; Wolfe & Smith, 2007). The item-level detail was particularly important to this project, as it provided information about how well students’ ratings of the difficulty of competencies matched theoretical expectations of when students likely develop these skills based on their age, gender, and other important characteristics. The Rasch item-person graphs (Figure 2; generated using Winsteps; Linacre, 2016), are an especially useful visualization tool that show how well the items (right side of each graph) are able to assess the full range of students’ self-reported competency (left side of each graph), and where additional items need to be written to fill in gaps where the scale is not able to assess all students’ social and emotional ability.

Figure 2 displays the item-person graph of the Relationship Skills (RS) subdimension between Phase One (2014) and Phase Three (2015) of the project, with the 27-item version from Phase One on the left panel and the 28-item version from Phase Three on the right panel. Markings on the left-hand side of each panel depict the number of students at each level of RS (each hashtag denotes 56 students; each dot denotes 1 to 55 students). Items (numbered top to bottom from most to least difficult) on the right of each panel reflect the range of latent relationship skill ability levels each item assesses. For 2015, the item numbering in the right panel also corresponds with the actual wording of the items in the Appendix. If the RS scale assessed all levels of student ability perfectly, items (right side) would be vertically distributed at equal intervals along the full spectrum of students' social emotional competency (left side). Further, if each item assessed a particular level of competency, there should also be minimal horizontal overlap of items (right side) indicating redundancy in targeting. For example, in 2014, items three, "I am able to describe my thoughts and feelings in ways that others understand" and four, "When I see classmates start to get upset, I can usually help them calm down" are on the same horizontal line, indicating they are both assessing the same ability level, and may be redundant. We created similar graphs for all subdimensions. Although in the interest of space, we only display graphs for Relationship Skills, we also discuss in the text features of the Self-Management of Emotions dimension, which had one of the highest ceiling effects in the 2013 version of the survey.

In addition to how well items can assess a range of SEC ability, category probability curves (Figure 3; 2014 depicted in left panel, 2015 depicted in right) were used to assess how reliably response options were utilized by students. For example, when students were presented with a scale that read "How true or untrue are these statements for you" with five response options, the category probability curve graph would show how reliably we could predict that students would select option one (*Never true for me*) versus option two (*Rarely true for me*). In a “perfect” version of this graph, each response option would “peak” at a different location above the others, which would indicate that all response options are well used.

The remainder of this article describes the three phases of the continuous measure improvement process (Figure 1) taken to address the primary statistical concerns of the measure and further interprets these category probability curve and item-person graphs. In Phase One, a large initial bank of items was developed collaboratively by the research-practice team, with items aligned to national and district practice standards. These items were tested using Rasch analyses to determine how well they captured a range of students’ SEC ability. In Phase Two, the research-practice team conducted latent class analyses and focus groups with students to better understand the statistical concerns (primarily the ceiling effect) testing revealed during Phase One. Finally, using knowledge garnered in Phase Two about the instrument and the larger context of survey administration in the district, the team conducted a second round of item revisions to improve the instrument in Phase Three, leading to a selection of both a 17- and 40-item set that was disseminated to other districts and used by WCSD. Throughout all phases, data-informed practices with educators and students were a central foundation of the continuous measure improvement approach, providing on-the-ground insights into data patterns and data needs of staff and students. It is hoped that the continuous measure improvement approach described can provide a framework for future research-practice teams to approach their measure development work.

Phase One: Initial Item Development

The partnership adopted CASEL's model of social and emotional learning, which includes five domains ("CASEL 5") of SEC: self-awareness, self-management, social awareness, relationship skills, and responsible decision-making (Payton et al., 2000). In order to create an open-source student self-report rating, the research-practice team had to develop a new set of items to correspond to these CASEL 5 competency clusters. As part of the CDI, WCSD had developed Social and Emotional Learning standards for each of the CASEL clusters for grade bands K-2, 3-5, 6-8, and 9-10, and 11-12. The standards provide educators with descriptions of the competencies at different developmental stages and sample classroom activities that support student growth in each area. Their purpose is to guide implementation and provide educators with key indicators to look for as students progress along a developmental pathway of SEC. For example, in early elementary grades, students should be able to "Distinguish among intensity levels of their emotions" as evidence of their self-management of emotions. To help teach this skill, staff might have students draw an "anger thermometer" and discuss how a person might feel physically as they move up the anger thermometer.

Using the CDI self-report measure as a starting point, the research-practice team both independently, and as a group, studied these SEL standards, writing new items to fill any gaps in coverage of the standards that existed. Several rounds of discussion and consensus occurred, with particular emphasis given to ensuring that high and low SEC ability were captured by the items, and that items assessed all SEC domains. Because the self-management and self-awareness standards encapsulated several different domains and merited more items to assess each component, these two scales were sub-divided into additional dimensions. Self-management was split into three dimensions: self-management of emotions, self-management of goals, and self-management of school work. Self-awareness was split into two dimensions: self-awareness of self-concept and self-awareness of emotions.

In 2014, a set of 112-items assessing eight subdimensions of SEC were developed, and included 27 different items that assessed Relationship Skills (RS), and 12 items that assessed Self-Management of Emotions (SME), specifically. Because the item bank was so large, the 112 items were randomly presented to students, so that each student only saw 40 total SEC items, but anchor items (presented to all students) allowed linking across students and to achievement measures. The large number of items was created to offer flexible use, for instance, to create a smaller scale representing the "best items" that captured the broadest range of SEC ability for future Climate Surveys in WCSD, and to study which items best predicted risk for dropout.

To increase the variability in student SEC captured by the instrument, the team also expanded the response options from a 4-point to a 5-point “Truth” scale that ranged from 1, *Never true for me* and 5, *Always true for me.* Finally, to inform whether a ceiling effect might have indicated that students were simply disengaging from the survey task, the SEC questions were counterbalanced within the larger Student Climate Survey, so that the entire SEC scale either appeared to students at the beginning of the survey or at the end of the survey. A more pronounced ceiling effect on SEC items placed at the end of the survey could indicate survey burnout.

In the spring of 2014, the 112-item measure was administered to students in grades 5, 6, 8, and 11. In grades 7 and 9, students continued to receive the original 28-item SEC measure to ensure longitudinal continuity for the larger CDI evaluation. Of the participating students in grades 5, 6, 8, and 11 (*n* = 7,618), 47% were Caucasian, 38% Hispanic, 5% Multi-Racial, 5% Asian, 2% Black, 1% American Indian or Alaska Native, and 1% Pacific Islander. Further, 11%

of these students had an Individualized Education Plan (IEP), 11% were English Language Learners (ELL), 46% qualified for free or reduced price lunch (FRL), and 49% were female. These demographics parallel WCSD's larger student population almost identically. The response rate on the SEC items was 79%.

Results and Implications

Although RS had one of the lower ceiling effects in 2013, we believed these items would resonate particularly well with adolescents during focus groups (see Appendix for item wording). We also describe the SME subdimension, as it had one of the highest ceiling effects in 2013, which substantially decreased as the measure developed. Returning to Figure 2 (left panel) to examine how well the RS items assessed different levels of ability, recall markings (hashtags and dots) on the left side of the graph display the number of students at each level of RS ability, with the items centered around a logit of zero. In 2014, student RS ability levels ranged from -4.39 logits (lowest level of ability) to 5.13 logits (highest level of ability). For the RS dimension, 7% of students rated themselves at the very highest level of ability, while only 1% rated themselves at the very lowest level of ability (Table 1 shows these metrics over the four years of the project). Although not graphed in the interest of space, student ability levels ranged from to -4.37 to 4.82 on our other example dimension, SME. In 2014, 7% of students also rated themselves at the very highest level of ability on SME and 1% rated themselves at the very lowest level of ability. Most students rated themselves with above average ability on most subdimensions, including RS shown in Figure 2. The right side of the 2014 graph displays each of the 27 items of the RS scale in order of difficulty (most difficult at the top, least difficult at the bottom, with each item placed at its average location on the 5 point scale). Relative to the student distribution, items were primarily assessing those with low-to-mid range SEC competencies. Even the most difficult item for the RS (Number 1 in the right side graph, “People look to me to solve conflicts”) was only able to assess students in the middle of the distribution of RS competency.

Neither the SME nor RS subdimensions had items that were able to effectively assess students at high levels of ability. Although the range of student RS ability was as high as 5.13, the most difficult item was targeted at 0.78. A similar, although slightly better picture emerged for SME (not shown in the graph). Student ability reached 4.82 logits, yet the most difficult item was targeted at 1.27, just slightly above average. Poor item targeting occurred at the low end for each scale as well. That is, RS items were also unable to assess students whose self-reported RS ability level was less than -0.84, where the least difficult item was targeted, though the range of student RS ability ranged as low as -4.39. SME items also were not well able to assess students with *lower* than average ability, with the easiest SME item only assessing students at -1.19 logits of ability, although the range of student ability floored at -4.37.

The response options also continued to show room for improvement in 2014. As noted above, Figure 3 displays the results of the category probability graph, with the left panel showing 2014 results. Of concern, the second lightest gray line reflects that the response probabilities for *Rarely true for me* (2’s) barely peaked, suggesting students had difficulty distinguishing this category from *Never true for me* (1’s) and *Sometimes true for me* (3’s) when selecting responses. Altogether, these findings initiated efforts to further understand the ceiling effect and why there was a lack of items able to assess students at the high end of ability. This included conducting a latent class analysis of the students who “maxed out” both the RS and SME subdimensions by self-reporting higher SEC abilities than items could assess and focus groups with students to understand how they approached and understood the survey, including its response options.

Phase Two: Item Evaluation and Exploration

In the second phase of the continuous measure improvement process, we used several strategies to identify which students were most likely to evidence high scores on individual item scales and to understand why the ceiling effect occurred.

Latent Class Analysis

Latent Class Analysis (LCA) was used to examine characteristics of the students who “maxed out” the SME subdimension (i.e., students who responded to all questions with the highest response option, *Always true for me*) and how they differed from students who did not “max out” the subdimension. The LCA was conducted using Stata 13 (StataCorp, 2013) using SME, which had a high proportion of students maxing out in 2013 (14%). In 2014, the total sample size for the LCA analysis was 5,652 students.[1] Of these students, 353 maxed out the SME scale.

In LCA, the researcher specifies several theoretically relevant categorical indicators (e.g. gender, reading levels, grade level) that might compose a particular “type” of person (e.g. 5th grade males who rate their social and emotional competencies poorly or 11th grade females who rate their competencies highly; Collins & Lanza, 2010). For this investigatory LCA analysis, we expected eight indicators of the survey structure and student characteristics would have an impact on how students approached and understood the survey items specified. The first was student level of risk for dropout in the previous (2013) school year. The second and third indicators reflected student growth percentiles (SGP) in reading and in math. The SGP is a measure of academic growth that is based on a comparison with other students with similar

[1] Of students who completed the SEC items in 2014, those who did not respond to all items administered and those who had missing data on the eight indicator variables were omitted from the LCA.

baseline standardized test scores (Betebenner, 2011). Students with lower growth (i.e., their growth was less than 60% of students with similar baseline scores) were indicated.

The fourth, fifth, and sixth indicators represent important demographic characteristics that may have influenced how students responded to the SEC items: gender, grade level, and English Language Learner status. The seventh and eighth indicators reflect characteristics of the items. One, as explained above, was SEC's placement in the overall Student Climate Survey, either first or last to help capture survey fatigue. The second was the conceptual orientation of the items. Students who received negatively oriented items but still responded with all 5s might have just been marking all items with the same response, and therefore, not have been fully engaged in the survey.

Model fit indices were examined for models specifying one to five latent classes, that is, different "types" of student and survey characteristics. Model selection was based on minimum Akaike information criteria (AIC) and the Bayesion information criteria (BIC) values which measure how well the data fit a specific model, with lower numbers indicating better model fit than higher numbers. The results of the LCA suggested either a two (AIC = 295.90, BIC = 361.63) or four (AIC = 268.94, BIC = 404.27) latent class solution best fit the data. The team focused on the four-class solution, which provided a more nuanced view of the patterns of students who "max out" the scale. The two LCA parameters reported here are gamma (γ) and rho (ρ). Gamma represents the percentage of the "maxed out" students belonging to the latent class group who maxed out the scale whereas rho represents the probability of having a particular characteristic within the latent class (Lanza, Collins, Lemmon, & Schafer, 2009). For comparison, we also present overall descriptive statistics for the full sample and "maxed" sample.

As Table 2 shows, there were four main groups of students who rated all of the items of the scale using the highest response option. Each of these groups may have had different motivations for this response pattern. The first group (termed “Disengaged”; γ = .11) was primarily boys (ρ = .22; gender = female) with higher reading (ρ = .97) and math (ρ = .85) SGPs who received the SEC items at the end (ρ = .13; items at beginning indicated). The second group (“Male lower comprehension”; γ = .36) was also primarily boys (ρ = .25) but with lower reading (ρ = .07) and math SGPs (ρ = .10) who received the SEC items at the end (ρ = .36). They were similar to the third group (“Female, lower comprehension”; γ = .13) who was mostly younger (ρ = .02; grade = reference is grade 8 or 11) girls (ρ = .62) who were at higher risk (ρ = .15; versus no or low risk), had lower reading (ρ = .14) and math (ρ = .06) SGPs and who received the SEC items at the end (ρ = .28). Groups two and three might not have understood the wording in some of the questions, and could have been guessing at their self-ratings as a result. This occurred especially when the SEC items were at the end of the survey, where poor item understanding might have led to disengagement in the task.

The fourth, and largest factor (“High achieving females;” γ = .40) consisted of mostly low-risk (ρ = .94) girls (ρ = .72) with higher reading (ρ = .46) and math (ρ = .53) SGPs who received similarly oriented (ρ = .96) SEC items at the beginning (ρ = .56). This last group of high-achieving females’ self-ratings did not vary when the SEC items were at the beginning or end of the survey, and they may have been students who really did believe they had high social and emotional competencies. Thus, the LCA provided some insight into various factors that may influence how students respond to surveys, which were further investigated through focus groups with students, the methodology and results of which are presented next.

Student Focus Groups

The research-practitioner team conducted focus groups with students to learn more about how the survey was administered across sites after LCA analyses pointed to several challenges with the way students may have been approaching and understanding the SEC survey. The purpose of the focus groups was to better understand: 1) the survey environment and directions provided to students taking the student climate survey; 2) students' understanding of the vocabulary and wording of items; 3) students' own impressions of what good or poor relationship competencies are; 4) students' perceptions of how interesting or boring the survey task was, and 5) whether other survey modes were more or less engaging. A total of nine focus groups were held with 74 students. All sessions were audio-recorded and transcribed verbatim. Students were recruited to participate by school staff (typically the principal or counselor) who were encouraged to randomly select students. Passive parental permission to participate in the focus groups was obtained, and students were provided pizza and snacks for their participation.

During the focus groups, students received a 9-item subset from the 27-item Relationship Skills subdimension from the 2014 version of the survey. This SEC subdimension was selected because students would have a broad range of examples of relationships at school and would easily generate discussion. In addition to the more specific activities and focus group questions tailored to each group, all students were asked questions about their survey environment (e.g. "what instructions were given?", "how comfortable did you feel responding?") and about their interpretation of what it means to have strong relationship skills ("how can you tell if someone your age has really good or poor relationship skills?"). Three types of groups were conducted that mirrored the LCA results, with tailored questions in each group: 1) elementary school students focused on comprehension of the items; 2) middle and high school groups focused on survey engagement; and 3) a high school group focused on new item generation. Two additional groups, one elementary and one middle, were held to identify a new response option structure.

Elementary school "Comprehension" focus groups. Three semi-structured, hour-long focus groups were held with 26 sixth grade students from three elementary schools (8-9 students at each site) to understand students' comprehension of items in the RS subdimension. Students received a survey with nine items from the RS item bank that mimicked the appearance of the actual survey. Students were first asked to rate themselves, then go back and identify words and items that were confusing or hard to understand. After responding to the mock items, students explained their comprehension of the items and why they found some to be difficult.

Middle and high school "Engagement" focus groups. A total of 18 high school (one all-male group, one all-female group), and nine middle school students (one mixed gender group) participated in focus group sessions to gather information about engagement in the survey and alternative options for collecting data. High school participants had previously dropped out of school and were in the process of returning to their traditional high school. Middle school participants came from one of the higher poverty schools in the district. Students in the "Engagement" focus group were asked to first take a nine-item survey in traditional paper-and-pencil format. Then they were asked to try a more hands-on version of the same survey, in which they received cards with the nine items and hand-sorted them into three cups with different category labels: 1, *Most Like Me (I am able to do them frequently)*; 2*, Somewhat Like Me (I am able to them some of the time)*; and 3*, Least Like Me (I am rarely able to them).* Then, students were asked a series of questions about which process was more engaging and why, which format provided a more accurate depiction of their skills, whether they noticed any patterns in the way they had sorted their items, and what ideas they had for making the survey format more engaging in the future.

High school leadership "Challenging Item Generation" group. Eight students from a high school leadership course participated in the "Challenging Item Generation" group. The item sorting activity described above was again used with this group, but students were asked to select one item each from their *most like me* and *least like me* category cups and describe why each item was more or less difficult for them and their peers to do. The purpose was to generate new items at the highest range of social and emotional competence that resonated with students' experience.

Elementary and middle school "Response Option" focus group. Finally, two focus groups with 13 students (one elementary and one middle school group) were held to identify a better response option structure. Students were given a randomized set of six items and asked to rate themselves on those items using four different, 5-point response option structures (i.e., *Not at all confident* to *Very confident*, *Not at all true* to *Very true*, *Poor* to *Excellent*, and *0-25% of the time* to *75-100% of the time*). They were then asked which they preferred, which most accurately assessed their SECs, and how they came to their answers using the different response options.

Results and Implications

Several major themes were revealed in the focus groups, many of which confirmed hypotheses about why the ceiling effect occurred, while other findings revealed a multitude of opportunities to improve the process of survey administration and the instrument itself. Elementary school students identified areas where the instrument could be made easier to understand, while middle and high school students discussed why the survey task was not always perceived as engaging to students. All students then contributed their ideas about what relationship skills mean to them, offering new phrases and suggestions for future items to include in the survey.

General survey perceptions. Students in all focus groups were asked to reflect about their reaction to participating in the Student Climate Survey. Overall, students were positive about the Student Climate Survey experience, believing that it was a unique opportunity to share their opinions, “It was fun, because you get to express yourself. ‘Cause not in every day you get to do that, or to give your own opinions.” Another student commented that the survey indicated school staff cared about their ideas and opinions, “It seems like when they have us do the survey, they actually want to know or actually care about what’s happening.” When asked about what proctors said to them during the administration, most students said that proctors had conveyed the importance of the survey and reminded students that the survey was confidential. Most students understood the survey as a tool used by staff to help improve the school, “They gather the ideas, the opinions of the students. Then they get an idea about what they need.”

Although many students were positive about the survey experience, some did not understand why the survey was administered and did not know that it was confidential. Several elementary and even middle school students referred to the survey as a “test”, with some worrying that they had answered incorrectly. One elementary school student, after a discussion about what the term “confidentiality” meant during the focus group explained, “No, they never told us it was confidential, so I thought if I did something wrong, then soon, it’d be posted on the news.” Another elementary school student said, “Some of my friends would have different answers. I was just worried that I had a totally different answer than them. I might be judged a little.” Several other students conveyed that teachers described the Climate Survey in negative terms, conveying that it was a distraction to learning and unimportant. As one high school student described:

> *When your teacher's like, 'All right, well, we have to go take this survey again for the hundredth time,' it's almost like a punishment. It takes like 15 minutes, you're just staring at a computer screen—every question is about the same thing.*

Students in high school also commented that despite having taken the survey several years in a row, they had never seen the results from it, nor had policies or practices ever seemed to change as a result of the survey. One high school student described his frustration with some of the questions on the Safety Survey this way:

> *You guys say that people take it every year, but it doesn't really make a difference. Like, when you ask questions about bullying, you think, 'Okay, then maybe the school's gonna do something about the bullying,' but every year you still take the same survey with the same questions, and nothing ever happens. I feel like just, after a while, people get tired of it, and it's like, 'Maybe I saw bullying, maybe I didn't—yeah, I did, it's not a big deal.'*

Other students commented that the questions often felt "too personal", which affected their willingness to be honest, "They're not gonna wanna give the right answer 'cause they think it's too personal. They'll wanna give false answers." In summary, although many students valued the opportunity to provide feedback about themselves and their schools, some students' concerns about confidentiality and beliefs that nothing ever changed as a result of their feedback limited their willingness to provide thoughtful, honest answers on the survey.

Understanding of relationship skills. Students across all focus groups were asked a series of questions about what relationship skills looked and felt like among students they knew. In general, students across the focus groups had similar beliefs about what it meant to have strong relationship skills, which mainly mirrored what was contained in WCSD's SEL standards. Many students noted that students with good relationship skills would talk to everyone, “They’re nice to people" or "They’re talking to people, not just in a certain group. They go over and talk to different groups.” Other students described good relationship skills in terms of attentiveness, saying that students with these skills, “Remember small details about what you said and bring them up later” or that they “Have good eye contact” or “Make you feel appreciated for talking to them.” Others commented that good relationship skills were evident when students were “mature”, “respectful of adults and authority”, and were “appropriate” when they should be (e.g. in class when the teacher was talking).

Persons exhibiting poorer relationship skills were characterized as being “two-faced” and “judgmental," would “make fun of other people”, were disloyal to friends, or were always “causing drama, wanting to be involved in other people’s business.” Students also believed people with poor relationships might be uncompromising, argumentative, “close-minded”, condescending or “have to be ‘right’ in any argument.” A few elementary school students believed that people with poor relationship skills would not take responsibility for mistakes, or would not do their school work. In sum, students generally agreed on what good and poor relationship skills were, and most often thought about their direct experiences with friends and classmates.

Comprehension of survey items. Most elementary school students across groups said the survey was “easy enough”, but others believed the survey was hard to understand. Students most frequently brought up colloquialisms like “agree to disagree” in the survey question “I can agree to disagree in an argument” as the most challenging terms to understand. Students often highlighted terms like “conflict”, “joint”, “resolve”, “project”, and “compromise" as challenging words. Students also struggled with double negatives (“There are very few people that I do not get along with at school”) and complex phrasings (“I am able to stand up for myself without putting others down”). Elementary school students also had a hard time responding to questions about events with which they had little experience (“I apologize when I learn that I upset a classmate”) or events that were too general (“I am comfortable joining a group when people are already doing something”). Finally, many students in elementary school struggled with the Likert-style response structure itself, with some having a hard time differentiating between similar response options like *Usually true for me* and *Always true for me*. One student suggested to her classmate who struggled to understand the difference, that in those instances, she could just pick a different option, *Somewhat true for me* because it was somewhat *usually* and somewhat *always*. These findings indicated that many students in elementary school struggled with the complex phrasing and vocabulary of the survey items.

Engagement of survey task. Middle and high school students collectively agreed that the hands-on item sorting activity was “more fun” and required them to cognitively engage with the items better than the traditional survey format, which they thought was boring and like many other tests they had to take at school. However, they believed the forced choice nature of the cup- sorting activity required them to rate themselves in ways they would not have if they were allowed to select any option. They also believed the focus group itself was a good way for educators to understand their competencies, and wished adults asked their opinions more often.

Student ratings of item challenge. Students in the middle and high school focus groups generally agreed on the items that were the easiest. Most believed that cooperating on school projects was one of the easier relationship competencies. They explained that school regularly required them to work in groups and they had developed strategies for working with team members over the course of their school careers. Students also indicated that getting along with others at school was easy for them to do, “We are around random people all the time at school, so this is something we do a lot.” Students in the high achieving, high school group overall rated themselves highly competent on most RS items. They noted that getting along with others at school was easy for them because they were “good at finding common interests” with most other people. They also believed that helping others solve conflicts was relatively easy, commenting that “friends trust my opinions.”

The groups differed more substantially on the items that were the most challenging for them which offered insights into new items to reduce the ceiling effect. Middle school students believed that describing their thoughts and feelings, understanding other people’s behaviors, and calming down other people were the most difficult for them. They said they often were unsure of what to say and did not want to make a situation worse or did not know how to empathize with something they had not gone through or did not fully understand. Students in the group that had dropped out of high school and recently re-engaged struggled the most with relationship competencies that dealt with healthy confrontation or argumentation, with one stating “yeah, I mean you pop off at me, I’m gonna pop off at you” and “people have to earn respect before I’ll be nice.” Students in the high achieving group had the most difficult time conceding arguments (“I’m stubborn”), saying that they preferred to persuade others to their opinions, especially because they felt they researched issues more than most and had better bases for their opinions than other people they knew, even teachers. They also had a harder time apologizing to others, saying that they did not like to admit when they were wrong, or that they felt uncomfortable engaging in confrontation. These conversations helped the team develop several new, more challenging items, including “Getting along with adults, even when we disagree”, “Helping others solve their disagreements", and “Introducing myself to a new student at school.”

Response options. Students reported liking the “Poor to Excellent” response structure the least of the four presented, saying that the options felt “judgmental.” Students liked the “Percentage of Time” response options the most of the four options, commenting that there was a difference between feeling confident that you can do something and actually doing it regularly. Students reported not liking the “Truth” response structure, as they believed it reflected attitudes about themselves more than the frequency they exhibited these skills. Taken together, the findings from student focus groups revealed several areas where item comprehension could be improved, and provided new language and ideas for additional items in the survey.

Phase Three: Item Refinement

Based on the results of the latent class analyses and student focus groups, two types of refinement efforts were undertaken to reduce the ceiling effect: 1) item and response structure revisions to improve comprehension and expand the item pool to assess higher SEC ability levels; and 2) improvements in the survey environment to ensure consistency and build buy-in for the importance of the survey among students and staff. During this final round of item revisions, an additional 26 items were written using student suggestions for new items and explanations of which were the most confusing items or phrases. To generate these new items and refine the previous year's item set, members of the research-practitioner team divided into pairs, one person from the research side and one from the practitioner side, to review each subdimension and exchange revised items back and forth. The full team reviewed the revised items to check again for readability and comprehension of the items, and utility of each for classroom-level instructional decision-making. The result was a bank of 138 items assessing eight competencies, with 28 items assessing relationship skills (see Appendix for the 2015, 28-item version of RS items).

The response option structure was changed from a 5-point “Truth” response option structure to a 4-point “Difficulty” response option structure with 1 (*Very difficult),* 2 (*Difficult),* 3 (*Easy*), and 4 (*Very easy).* Although not tested in focus groups with students, the team believed this response structure was non-judgmental (unlike the “Poor/Excellent” version), and used simpler language than the "Truth" response option structure. It also paralleled what students said they liked about the "Percentage of time" scale, in that it better reflects ease of using these skills, rather than simply attitudes about themselves like the “Truth” response structure, but was less cognitively demanding than the “Percentage of time” scale was for the survey environment.

Additionally, a new response formatting structure was tested with half of all students. Qualtrics offers several different options for response structures besides a traditional multiple choice format. Several of these options were tested during a focus group with students. In this context, students responded the most favorably to being able to slide a bar to their preferred response option on an implicitly continuous scale. We anticipated that this more engaging response structure might reduce the ceiling effect for two reasons. First, it might better match students’ preferences for the more fun, hands-on activity like the cup sort used in focus groups. Second, it might enhance attention as it differed from response structures they saw elsewhere in the survey (and in prior years). As Table 1 displays, however, this new “slider” option actually substantially *increased* the proportion of students who “maxed out” the subdimension by selecting the highest response option (*Very Easy*) on all items. In fact, for the RS subdimension, the slider function produced a max-out rate of 16%, double the max-out rate found on the traditional Likert-style response format that same year (8%). These findings were paralleled on the SME items, with a max-out rate of 8% when students used the slider function compared with 5% when they used the Likert-style response format. As a result, in Figures 2 and 3, only results from the traditional Likert-style response format are shown for comparability across years.

The same Rasch analyses from Phase One were used to assess whether the revisions reduced the ceiling effect and helped measure a broader range of student ability (see Table 1 for response rates, number of items in each subdimension, and key indicators of ceiling and floor effect across years). The right panel of Figure 2 displays the Rasch item-person graphs for the 2015 RS subdimension version. For the 28 RS items in Phase Three (2015), the range of SEC covered by the average item locations almost doubled to 3.25 (-1.84 to 1.41) compared to 1.62 in Phase One (-0.84 to .78). In fact, whereas in the previous year, the most difficult item could not even assess students of average SEC ability, in 2015, the most difficult item ("Sharing what I am feeling with others") was able to assess students with above average ability levels. At the same time that item spread improved, the ceiling effect remained nearly the same (8% in 2015 vs. 7% in 2014; note that the right-hand panel only reflects data from the Likert-style response format, about half as many students in 2015 as in 2014, which explains the fewer hashtags at the top of the graph in 2015, despite similar ceiling effects.). Although not shown because of space limitations, the most difficult SME items in both Phase Three ("Concentrating when there is a lot of noise around me") and Phase One ("When I get upset I can't concentrate") targeted students only slightly above average ability. Although the SME subdimension targeted a higher level of ability than previously, the overall range of ability the items assessed did decrease slightly in Phase Three to 2.28 (-1.51 to 0.77) from 2.46 (-1.19 to 1.27) in Phase One. Altogether, the better targeting of students with above average ability corresponded to the percentage of students who "maxed out" the SME scale declining somewhat, from 7% in 2014 to 5% in 2015.

Analyses of the both the RS and SME dimensions also showed considerable improvement in how the students were using the new response option structure. In the right-hand panel of Figure 3, all response options clearly “peak” in 2015, whereas one response option just barely peaked in 2014 (left panel), indicating that students were better able to distinguish among response options with the 4-point “Difficulty” structure than with the 5-point “Truth” response option structure. Taken together, results of the 2015 Rasch analysis demonstrated that the response options were better used, that the ceiling effect improved for some dimensions, and that the RS items were able to assess a broader range of student SEC ability. That said, although there were fewer students whose abilities exceeded the instrument's capacity to assess their RS and SME competencies, there remained a deficit of items able to assess students at the highest range of RS and SME ability levels. Continued progress on the ceiling effect is discussed in the next section and the conclusion.

Implications

Although the measure had room for improvement and there were plans for future iterations of refinement, the team felt the 2015 version had strong enough properties for its intended purpose (see Washoe County School District & Collaborative for Academic, Social, and Emotional Learning, 2016). In 2016, the final year of the project, the team focused on selecting shorter form sets of items from the bank of 138 items. The research-practitioner team spent considerable time selecting a 40- and 17-item version from the larger set. Items were chosen based on the Rasch difficult ratings, with preference for items that would cover the full range of student ability. Items were also selected based on their theoretical centrality to the CASEL 5 domains. For the 40-item version, a total of six RS items were selected from the 138 item bank to represent the RS dimension in the 40-item version and four items were selected to represent the SME dimension. In 2016, all students received this identical 40-item version, which included the short-form 17-item measure assessing overall SEC competency tested in 2015 (though further refined in 2016) embedded within it. The “slider” response format was removed, and all students received the Likert-style response option format at the beginning of the survey in 2016.

The six-item Relationship Skills subdimension from the 40-item instrument again showed a diminished ceiling effect and assessed a broad range of RS ability in comparison to the Phase One survey. Although the range of item locations decreased somewhat to 2.11 logits (the easiest item had a Rasch location of -0.94 whereas the most difficult item had a Rasch location of 1.17), more compellingly, there was virtually no ceiling effect, as the number of students who "maxed out" the subdimension was dramatically reduced (3% in 2016 compared with 8% in 2015). For SME items, the ceiling effect was even further reduced over time. In 2014, 7% of students “maxed out” the subdimension, compared with just 5% in 2015, and 3% in 2016. These findings show that even though the shortened version of the instrument had room for continued refinement, it reflected meaningful improvement in its psychometric properties over previous years.

In 2016, the research-practitioner team also explored the relationship between the SEC measure and student outcomes. Preliminary findings on the 17-item, composite SEC measure indicate that students’ SECs were related to several academic and behavioral outcomes. These findings were shared through a series of presentations to WCSD educators and participants at the 2016 CDI event in Reno, NV (Washoe County School District & Collaboration for Academic and Social Emotional Learning, 2016, February). Using a series of multilevel regression analyses, students’ self-reported SECs on the 17-item instrument were compared against their standardized test scores in reading and math (grades 5, 6, and 8), weighted GPA (grade 11), and their number of days absent that same year. Multilevel logit analyses examined the association between SECs and whether students had been suspended at least one time the year they completed the survey. In all analyses, we examined these associations while controlling for grade level, gender, IEP status, LEP status, FRL status and outcome data from the year prior. The Early Warning Risk Index categorization was only included as a covariate when examining standardized test scores and GPA because of its high correlation with student absenteeism and suspension.

These regression analyses revealed that students' self-reported SECs were positively related to their standardized reading scores, *b* = 7.91, *p* < .001, standardized math scores, *b* = 8.19, *p* < .001, and weighted GPA, *b* = 0.18, *p* < .001, and negatively related to the number of days they were absent, *b* = -0.07, *p* < .001, and likelihood of being suspended that year, *b* = -0.33, *p* = .01. Results from these analyses were shared with school staff during SEL trainings to highlight the relationship between SECs and their link to positive academic and behavioral outcomes. Presenting local data demonstrating a positive relationship between social and emotional skills and student academic and behavioral outcomes helped build buy-in for SEL programming among school staff, highlighting the importance and potential impact of educators' hard work.

All Phases: Data-Driven Practice

In addition to item revisions, efforts were undertaken to address the issues raised through the LCA and focus group analyses, including improving the survey environment, increasing staff investment in the importance of the survey, and improving consistency in the directions provided to students during administration. A proctoring video with instructions for the survey was developed and disseminated to school staff prior to survey administration (adapted from Bradshaw, 2014). Students in the video described why the survey was being administered and what would be asked within the survey, their rights as student participants to opt out of the survey or skip questions that made them uncomfortable, who would see the results of the survey, and where the reports would be posted on the district website if they wanted to see the results of their school's survey (Washoe County School District, 2016).

Additionally, WCSD began sharing the results from the focus groups and LCA with educators to increase buy-in for the survey prior to survey administration. WCSD's Office of Accountability developed a training in collaboration with the SEL Department that highlighted the importance of a well-controlled survey environment and also encouraged educators to discuss the results of student focus group findings with their students and colleagues after dissemination (for more information, see Schamberg et al., in press). The LCA and focus group results were presented to district leaders and community members as part of a larger conference about district data, and to hundreds of school staff during SEL implementation trainings.

Findings from the latent class analyses and student focus groups also helped transform the way in which WCSD approached student voice and school improvement processes. Though it should have come as no surprise that students would be interested in the results of the Student Climate Survey and the other academic and behavioral measures that reflect their educational experiences, the focus groups highlighted the notable absence of student voice in district data-based decision-making processes. To begin to address this concern, the WCSD Office of Accountability held its first Data Symposium for students in 2015, so that students would have an opportunity to reflect on climate survey results and offer input and recommendations based on the data. High school students presented all of the data to fellow students, helping them build capacity to facilitate conversations around district educational data (e.g. academic achievement gaps, office discipline referral data, and social and emotional competencies). The district has since hosted this event annually, which includes a full day focused around exposing students to district data and engaging them in school improvement decision-making with their school leaders.

Further, the array of item locations on all SEC dimensions were shared with staff during trainings and local conferences so that educators could reflect on the conceptual meaning of the empirical item orders, and discuss ways in which the policies and practices of schools and the district might influence what skills students found the easiest or hardest. For example, using the RS item locations, many staff noticed that items rated as least difficult by students were also those that were taught more frequently in school (e.g., getting along with students and staff, working in teams for class projects), and those which were more difficult for students were not regularly taught or encouraged (e.g. sharing feelings with others). In addition to direct professional development, all Climate and Safety Survey data was shared with each school at the start of the year as part of their annual data profile, a large booklet of data that included academic data, behavior and attendance data, and other climate survey results. The WCSD Office of Accountability held workshops and small group sessions to help leadership teams review data and encourage thoughtful examinations and conversations around the SEC data in relation to other important measures of academics and school climate and culture. For example, student self-reports that managing their emotions when they are upset was one of the most difficult SEC items on the SME instrument might have mirrored increases in office behavioral referrals that same year.

As a result of these efforts, Climate Survey, and particularly SEC data, is now one of the primary sources of data school leadership use to identify areas of need at their school in their annual School Performance Plans. Schools use the data to identify specific student competencies on which to focus Social and Emotional Learning programming as well as to evaluate new programs and initiatives designed to improve areas of need. Several school administrators have even begun sharing SEC and Climate Survey data with their students using the data-based decision-making strategies modeled at the Annual Student Data Symposium. Finally, because the instruments are open-source, school staff also have the SEC instrument outside of the formal Climate Survey administration in their own school improvement practices. A number of schools use the 17- and 40-item instruments for ongoing progress monitoring between Climate Survey administrations. WCSD’s SEL Department developed an activity in which school staff complete portions of the SEC item bank about themselves so that they can self-assess their own capacity to teach students these skills. The open-source nature of these instruments has led to several innovations in how the SEC measures are used and how the data is shared with both educators and students.

Conclusions

The IES Research-Practitioner project and the continuous measure improvement approach described in this paper allowed for an iterative, longitudinal revision process for a social and emotional assessment focused on measuring students' self-reported social and emotional competencies across a range of abilities. By expanding the measure to assess a range of skills across competencies, simplifying language, and improving the consistency of the survey environment, the ceiling effect was reduced and the measure can now capture sufficient variation in student SEC ability so the team feels more confident in its use in regression models predicting student risk for dropout, one of the partnership's intended uses for the measure. As an added benefit, the approach taken ensured that the instrument aligned to a strong theoretical model of social and emotional competency developed by CASEL as well as to local SEL standards used by practitioners in WCSD. This alignment ensures that the instrument used language that parallels what educators hear during professional development, and what educators use to guide implementation and progress monitoring of Social and Emotional Learning in the district.

That said, although the ceiling effect improved considerably over the four years of development, there remains a need to further develop items to assess students at the highest level of social and emotional ability levels. Our work also reinforces general challenges associated with self-report measures of social and emotional competencies. Self-report responses, especially student responses (La Greca, 1990; Miller, 2012), may be heavily influenced by social desirability. This is particularly true when inconsistencies in survey environments exist. Students in our focus groups indicated not taking the questionnaire very seriously when proctors did not convey its importance, and not responding honestly when they questioned the confidentiality of the survey. These administration challenges are likely not unique to WCSD, but may parallel the way in which non-cognitive measures are administered in districts across the country. Certainly the WCSD team learned how difficult it was to ensure private, consistent survey environments for the 22,000 students who take the survey each year. These survey concerns raise important questions about using this and other mass-administered measures of social and emotional competence in high-stakes contexts different from our partnership uses, like educator evaluations and school accountability.

From the outset, the team focused on building an instrument, and accompanying data dashboards, that could be used to help school staff drive school improvement and instructional decision-making. One of the remarkable outcomes of the research-practice partnership was the knowledge transfer that occurred between psychometric and substantive experts, leading to several innovations in the way the data was used and disseminated throughout the project. All members of the team learned how to use new psychometric tools, like the item-person graphs (Figure 2), so that by the end of the project, practitioners and researchers alike understood these tools in new and deeper ways. Practitioners were committed to disseminating not only the survey data, but findings from the research project itself to school staff, district leadership, and students. They subsequently developed innovative strategies for displaying these complex statistical tools for educators and students, who had incredible insights into the data patterns that existed and provided feedback about their greatest needs for a social and emotional measure. The intended use of the measure was to help educators identify student needs and supports along the path to graduation and assess aggregate needs across schools and grade levels. To this end, the partnership worked deliberately to ensure that educators and students were key partners in the development of the tool throughout the process, and that the resulting instrument met their needs first.

The amplified discussion on social emotional learning and non-cognitive factors in public education today, coupled with the continuing movement for data-driven decisions, calls for deliberate work to be done in the realm of SEC measurement and data use. The initial yields of our research-practitioner project, including a 17-item short-form, 40-item long-form, and 138-item bank of self-report instruments that are open source and flexible will add some value in this space. Perhaps even more meaningful, however, were the lessons learned throughout the process from several rounds of analysis, collaboration, iteration, refinement, and student voice. We progressed from originally being eager to analyze self-report SEC data and its relation to academic and behavioral outcomes, to realizing those data did not initially provide reliable signals due to several factors, including its psychometric properties, response option structure, and survey environment, and finally to a point where the data, educators, and students all agreed the measurements were yielding meaningful and useful enough information to begin use in regression models and decision-making. Along this path, we also learned the importance of making the measurement project understandable and relevant to educators implementing social and emotional learning each day in their classrooms.

The current partnership and project is ongoing, and some progress remains unfinished. More work needs to be done to assess the measure's relationship to academic outcomes, and the partnership is continuing such efforts including by testing potential moderating effects between SECs and risk factors embedded in WCSD's Early Warning Risk Index System. We are also planning future projects which may examine the extent to which the current measure captures growth in SECs after school-wide and individual SEL interventions as well as the extent to which the student self-report items show concurrent validity with alternative measures of SEC (e.g., validated but longer and more costly measures). This continued work will be helpful for practitioners to the extent that it provides information on the malleability of SEC and the impact those competencies have on moderating risk for students not graduating. In order for practitioners to improve and support the systemic development of social and emotional competencies among the nation's students, they will need a useful appraisal of those competencies through time and through developmental stages of students they serve. The partnership described here illustrates not necessarily the development of a perfectly reliable and validated assessment, but a process by which iterative research, practice, and listening can lead to improved measures along with refined understanding by all stakeholders of the wide domain they seek to assess.

References

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education (2014). *Standards for educational and psychological testing*. Washington, DC: American Educational Research Association.

American Institutes for Research & Collaborative for Academic, Social, and Emotional Learning (2013). *Student self-report of social and emotional competencies.* Washington, DC and Chicago, IL: Authors.

Arsenio, W. F., Adams, E., & Gold, J. (2009). Social information processing, moral reasoning, and emotion attributions: Relations with adolescents’ reactive and proactive aggression. *Child Development, 80*, 1739–1755.

Balfanz, R., & Byrnes, V. (2006). Closing the mathematics achievement gap in high-poverty middle schools: Enablers and constraints. *Journal of Education for Students Placed at Risk (JESPAR)*, *11*, 143-159.

Balfanz, R., Herzog, L., & Mac Iver, D. J. (2007). Preventing student disengagement and keeping students on the graduation path in urban middle-grades schools: Early identification and effective interventions. *Educational Psychologist*, *42*, 223-235.

Barry, M., & Reschly, A. L. (2012). Longitudinal predictors of high school completion. *School Psychology Quarterly*, *27*(2), 74–84.

Betebenner, D. W. (2011). *A technical overview of the student growth percentile methodology: Student growth percentiles and percentile growth projections/trajectories*. Dover, NH: National Center for the Improvement of Educational Assessment.

Bradshaw, C. (2014, November). *Measurement systems to assess individual- and population-level change.* Presentation in *Innovations in design and utilization of measurement* *systems to promote children's cognitive, affective, and behavioral health.* Workshop conducted by the Institute of Medicine and National Research Council, Washington, DC.

Coburn, C. E., Penuel, W. R., & Geil, K. E. (2013). *Research-practice partnerships: A strategy for leveraging research for educational improvement in school districts.* New York, NY: William T. Grant Foundation.

Collins, L. M., & Lanza, S. T. (2010). *Latent class and latent transition analysis: With applications in the social, behavioral, and health sciences.* Hoboken, NJ: Wiley and Sons.

Denham, S. (2016). Assessment of SEL in educational contexts. In J. A. Durlak, C. E. Domitrovich, R. P. Weissberg, & T. P. Gulotta (Eds.), *Handbook of social and emotional learning: Research and practice* (pp. 285-300). New York, NY: Guilford Press.

Domitrovich, C. E., Durlak, J., Staley, K., & Weissberg, R. P. (provisional acceptance). Social-emotional competence: An essential factor for promoting positive adjustment and reducing risk in school children. *Child Development*.

Farrington, C. A., Roderick, M., Allensworth, E., Nagoaka, J., Keyes, T. S., Johnson, D. W., & Beechum, N. O. (2012). *Teaching adolescents to become learners. The role of non-cognitive factors in shaping school performance: A critical literature review.* Chicago, IL: University of Chicago Consortium on Chicago School Research.

Gordon, R. A. (2014, November). *Assuring high quality in publicly funded child care and preschool: A cautionary tale.* Presentation in *Innovations in design and utilization of measurement systems to promote children's cognitive, affective, and behavioral health.* Workshop conducted by the Institute of Medicine and National Research Council, Washington, DC.

Gordon, R. A. (2015). Measuring constructs in family science: How can IRT improve precision and validity? *Journal of Marriage and Family, 77*, 147-176.

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*, 2283–2290.

La Greca, A. M. (1990). *Through the eyes of the child: Obtaining self-reports from children and adolescents.* Needham Heights, MA: Allyn & Bacon.

Lanza, S. T., Collins, L. M., Lemmon, D. R., & Schafer, J. L. (2009). PROC LCA: A SAS procedure for Latent Class Analysis. *Structural Equation modeling, 14*, 641-694.

Linacre, J. M. (2016). *Winsteps® Rasch measurement computer program*. Beaverton, OR: Winsteps.com.

McKown, C. (2016). Challenges and opportunities in the direct assessment of children's social and emotional comprehension. In J. A. Durlak, C. E. Domitrovich, R. P. Weissberg, & T. P. Gulotta (Eds.), *Handbook of social and emotional learning: Research and practice* (pp. 320-335). New York, NY: Guilford Press.

Miller, A. L. (2012). Investigating social desirability bias in student self-report surveys. *Educational Research Quarterly, 36*, 30-47.

Payton, J. W, Wardlaw, D. M., Graczyk, P. A., Bloodworth, M., Tompsett, C. J., & Weissberg, R. P. (2000). Social and emotional learning: A framework for promoting mental health and reducing risk behavior in children and youth. *Journal of School Health, 70*, 179-185.

Qualtrics (2015). *Qualtrics*. Provo, Utah: Qualtrics. Retrieved from http://www.qualtrics.com.

Schamberg, R. S., Domitrovich, C. E., Davidson, L. A., Hayes, B. I., Shaffer, T., Gordon, R. A., …Weissberg, R. P. (in press). The Collaborative for Academic, Social, and Emotional

Learning (CASEL) and Washoe County School District (WCSD) social and emotional learning assessment partnership. *Research-policymaker collaboration: Strategies for launching and sustaining successful partnerships*. New York, NY: Taylor and Francis.

StataCorp (2013). *Stata Statistical Software: Release 13*. College Station, TX: StataCorp LP.

Tseng, V. (2012). The uses of research in policy and practice. *Social Policy Report, 26*(2), 2-16.

Valiente, C., Swanson, J., & Eisenberg, N. (2012). Linking students' emotions and academic achievement: When and why emotions matter. *Child Development Perspectives*, *6*(2), 129–135.

Washoe County School District. (2016, April 2). *Climate Survey Invitation.* Retrieved from https://www.youtube.com/watch?v=mQGWlQj87DM&feature=youtu.be

Washoe County School District & Collaborative for Academic, Social, and Emotional Learning (2016, February). *Assessing to serve students: A progress report on WCSD & CASEL's research-practitioner grant*. Presentation at the Cross-Districts Learning Initiative Event, Reno, NV.

Wild, C. L. & Ramaswamy, R. (2008). *Improving testing: Applying process tools and techniques to assure quality*. New York, NY: Lawrence Erlbaum Associates.

Wolfe, E. W., & Smith, E. V., Jr. (2007). Instrument development tools and activities for measure validation using Rasch models: Part II-validation activities. In E. V. Smith, Jr. & R. M. Smith (Eds.), *Rasch measurement: Advanced and specialized applications* (pp. 243-290). Maple Grove, MN: Journal of Applied Measurement Press.

Table 1.

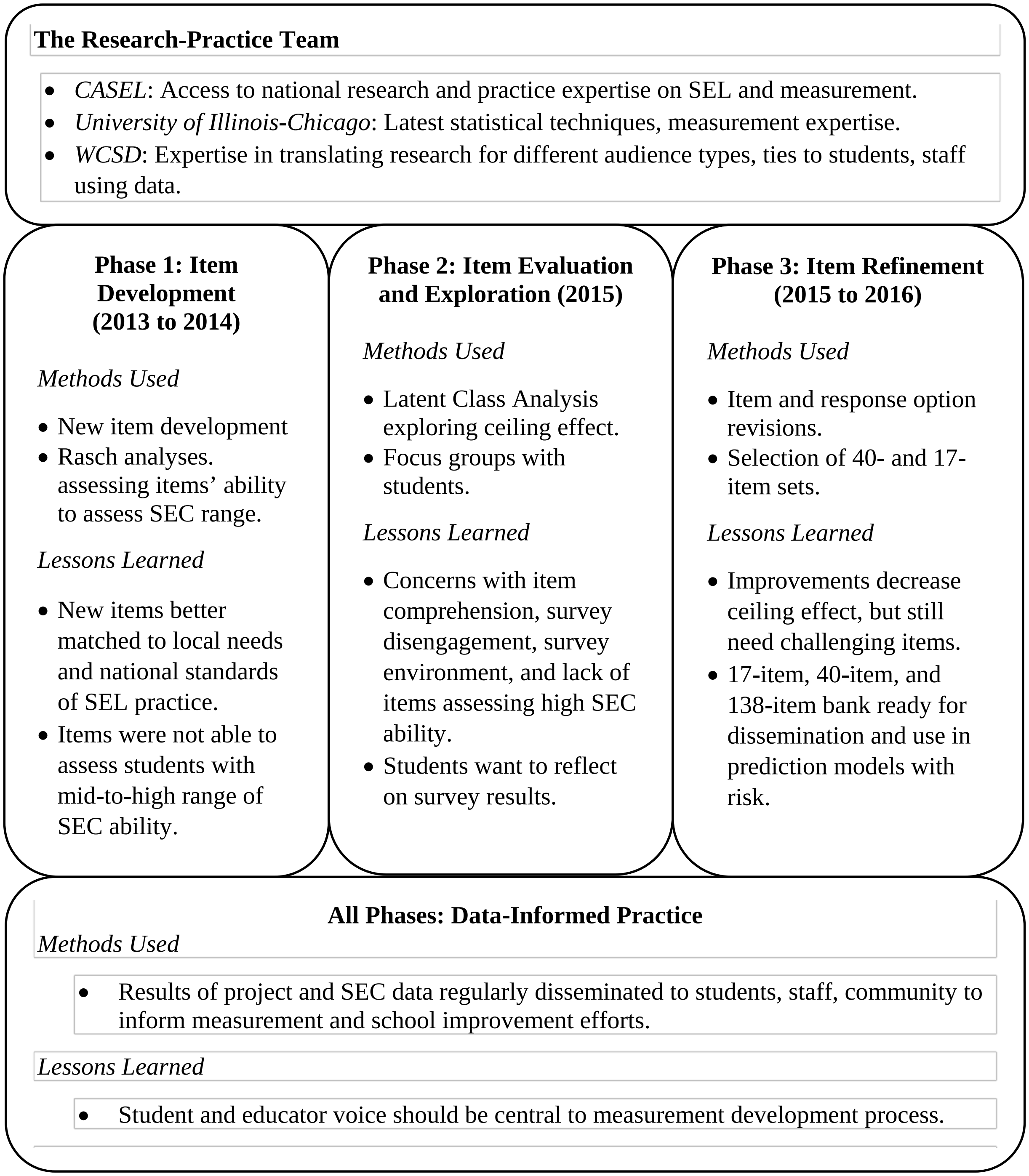
Description of full SEC instrument and Relationship Skill (RS) and Self-Management of Emotions (SME) competency subdimensions over time.

	Year One	*Year Two*	*Year Three*	*Year Four*
	2013	*2014*	*2015*	*2016*
			Likert *(Slider)*	
Total SEC Items	28	112	138	40
# of Responses	6,554	7,618	3,735 (*3,690*)	7,490
# of Dimensions	5	8	8	8
Response Rate	79%	79%	80%	81%
RS Items Only	6	27	28	6
% at Highest SEC Level[1]	10%	7%	8% (*16%*)	3%
% at Lowest SEC Level[1]	1%	1%	1% (*1%*)	0.4%
SME Items Only[2]	7	12	22	4
% at Highest SEC Level[1]	14%	7%	5% (*8%*)	3%
% at Lowest SEC Level[1]	1%	1%	1% (*.3%)*	1%

[1] Students who rated themselves at the highest/lowest level of ability the instrument could assess.
[2] In 2013, self-management was a single dimension, whereas in 2014 and all years following it was comprised of three dimensions, including self-management of emotions.

Figure 1.

CASEL-WCSD continuous measure improvement approach to building and evaluating an SEC measure that can assess full range of student ability levels.



The Research-Practice Team

- *CASEL*: Access to national research and practice expertise on SEL and measurement.
- *University of Illinois-Chicago*: Latest statistical techniques, measurement expertise.
- *WCSD*: Expertise in translating research for different audience types, ties to students, staff using data.

Phase 1: Item Development (2013 to 2014)

Methods Used

- New item development
- Rasch analyses. assessing items' ability to assess SEC range.

Lessons Learned

- New items better matched to local needs and national standards of SEL practice.
- Items were not able to assess students with mid-to-high range of SEC ability.

Phase 2: Item Evaluation and Exploration (2015)

Methods Used

- Latent Class Analysis exploring ceiling effect.
- Focus groups with students.

Lessons Learned

- Concerns with item comprehension, survey disengagement, survey environment, and lack of items assessing high SEC ability.
- Students want to reflect on survey results.

Phase 3: Item Refinement (2015 to 2016)

Methods Used

- Item and response option revisions.
- Selection of 40- and 17-item sets.

Lessons Learned

- Improvements decrease ceiling effect, but still need challenging items.
- 17-item, 40-item, and 138-item bank ready for dissemination and use in prediction models with risk.

All Phases: Data-Informed Practice

Methods Used

- Results of project and SEC data regularly disseminated to students, staff, community to inform measurement and school improvement efforts.

Lessons Learned

- Student and educator voice should be central to measurement development process.

Figure 2.

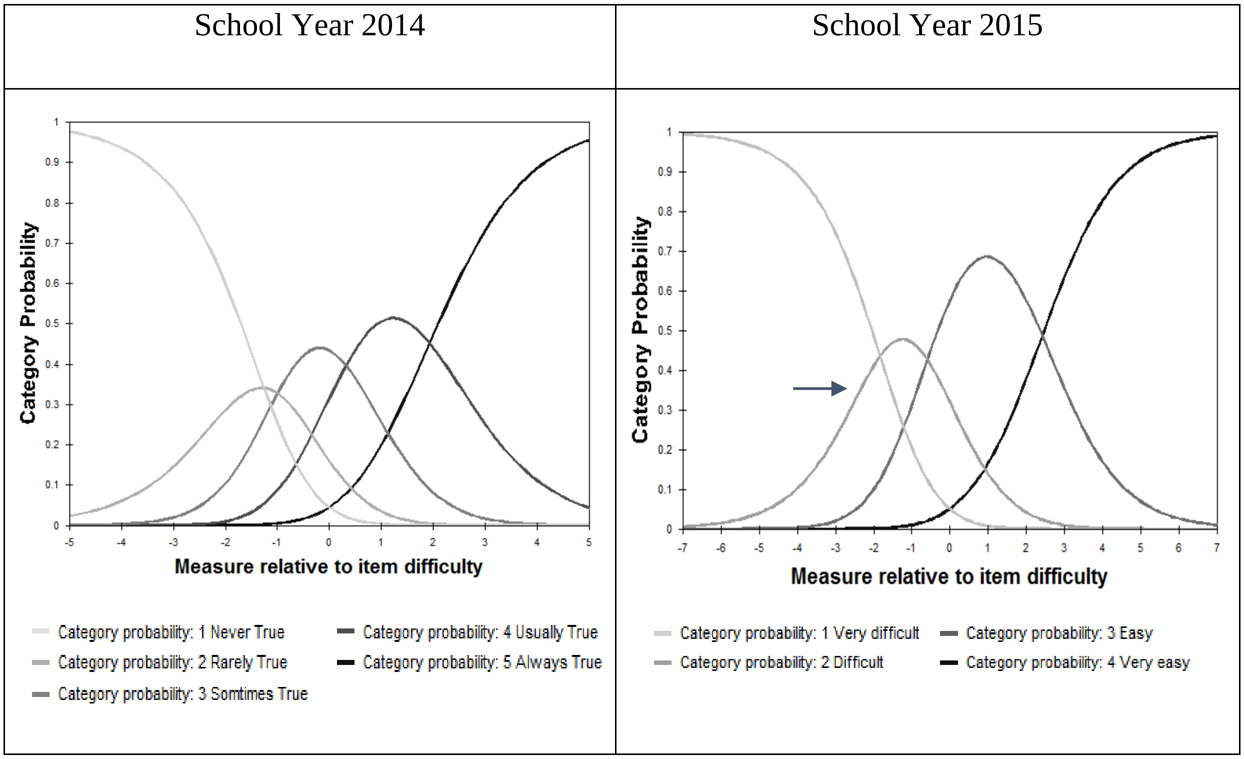
Rasch item-person graph of Relationship Skill subscale from 2014 (left) to 2015 (right).

Logit Unit	Person	Item	Person	Item
4	.#########		.#####	
	.		.	
	.		.	
	.##		.	
	.##		.	
	.		.	
	.		.	
3	.		.	
	.#		.	
	.###		.#	
	.###		.#	
	.#		.	
	.#		.	
	.####		.#	
2	.###		.#	
	.##		.#	
	.####		.#	
	.####		.#	
	.#####		.#	1
	.######		.#	2
	.#####		.#	
1	.######		.##	
	.#####		.##	3 4
	.#####	1	.###	5
	.#####	2	.##	
	.#####	3 4	.##	6
	.#####	5 6 7 8 9	.#	7 8 9 10 11
	.####	10 11 12 13 14	.##	12 13 14
0	.#####	15 16	.##	15
	.#####	17	.##	16 17 18 19 20
	.####	18 19 20 21 22	.#	21
	.###	23 24 25	.#	22
	.##	26	.#	23 24
	.#		.#	
	.#	27	.#	25
-1	.#		.	26
	.#		.	
	.		.	27
	.		.	
	.		.	
	.		.	
	.		.	28
-2	.		.	
	.		.	
	.		.	
	.		.	
	.		.	
	.		.	
	.		.	
-3	.		.	
	.		.	
			.	
			.	
			.	
-4	.		.	

Each '#' is 56 students. Each '.' is 1 to 55 students. Person ability and item locations simultaneously estimated on the latent RS dimension in logit units along the vertical axis, with higher numbers indicating higher ability. Items numbered sequentially for each year (1 = most difficult item), although content varied (e.g., Item 1 in 2014 does not necessarily reflect the same content as Item 1 in 2015). Item numbers in 2015 correspond with item numbers and wording in Appendix. In 2014, *n* = 7,618 (7% "max out"). In 2015, *n* = 3,735 (8% "max out").

Figure 3.

Category probability curves showing students' use of five response options ("Truth response structure") on the 2014 Relationship Skills subdimension versus four response options ("Difficulty structure") in 2015.[1]



[1] *Note*: 2015 chart only uses data from the Likert-style response option survey format.

Table 2.

Four-class LCA solution of students who "maxed out" SEC items in 2014.

	Overall Descriptives (*n* = 5,652)	Maxed Sample (*n* = 353)	Four Classes			
			1	2	3	4
Latent Class Prevalences (γ)			.11	.36	.13	.40
Item Response Probabilities (ρ)						
Risk Category = No or Low	.70	.69	.67	.61	.15	.94
Reading SGP > 60	.32	.33	.97	.07	.14	.46
Math SGP > 60	.31	.35	.85	.10	.06	.53
Gender = Female	.49	.48	.22	.25	.62	.72
Grade = 8 and 11	.48	.48	.51	.61	.02	.51
Non-ELL	.90	.86	.83	.88	.48	.99
Positive valenced items only	.52	.82	.70	.76	.66	.96
Items at beginning	.50	.41	.13	.36	.28	.56

Notes: SGP = Student Growth Percentile; ELL = English Language Learner

Appendix

Items in 2015 and 2016 Relationship Skills Subdimension

Question stem: How easy or difficult is each of the following for you?

1. Sharing what I am feeling with others.[*]
2. Joining a group I don't usually sit with at lunch.
3. Talking to my friends about how I feel when I am upset with them.
4. Talking to an adult when I have problems at school.[*]
5. Joining a group that is already talking.
6. Talking to classmates about why they feel a certain way.
7. Forgiving classmates when they upset me.
8. Forgiving myself if I hurt someone's feelings, after I apologize to them.
9. Working out disagreements on group projects.
10. Getting along with others even when I am having a bad day.
11. Helping other people solve their disagreements.
12. Getting along with adults at school even when we disagree.
13. Stopping myself before I hurt someone's feelings.
14. Getting along with classmates even if I disagree with them.
15. Fixing problems I am having with my friends.
16. Introducing myself to a new student at school.[+]
17. Using my skills to make my group better.
18. Helping classmates calm down if they're upset.
19. Making friends with people who have different opinions than me.
20. Getting along well with anyone my teacher assigns me to work with.
21. Making sure that everyone's ideas are heard in a group.
22. Forgiving classmates when they apologize to me.
23. Apologizing if I ever upset a classmate.
24. Respecting a classmate's opinions during a disagreement.[*]
25. Getting along with my classmates.[*]
26. Being polite to classmates.
27. Getting along with my teachers.[*]
28. Being polite to adults.

Items ordered from most (top) to least (bottom) difficult for students.

[*] Item used for 2016 version.

[+] Item reworded for 2016 to, “Being welcoming to someone I don’t usually eat lunch with.”

Response options for 2015 and 2016 ranged from 1, *Very difficult* to 4, V*ery easy*.