Using Social and Emotional Competency Measures for Research, Evaluation, and School Improvement

James Soland and Laura S. Hamilton
Early Education
Miscellaneous

Introduction

Schools that serve K–12 students in the United States are responsible for preparing graduates to succeed in higher education, the workplace, and civic and community life. Success in each of these realms requires a broad set of competencies (skills, knowledge, and dispositions/mindsets) applicable across disciplinary or academic domains. For instance, surveys of employers and analyses of job postings suggest that cross-cutting competencies such as communication and critical thinking are in high demand.[1] A 2012 National Research Council report[2], "Education for Life and Work," presented a framework comprising three domains of competencies: cognitive, interpersonal, and intrapersonal. Since that report was released, numerous frameworks have been developed to organize and define cross-cutting competencies,[3] which educators and researchers describe with a variety of adjectives such as "21st-century," "durable," "soft," "transferable," and "future-ready." K–12 educators and people who train or support them often use the phrase "social and emotional competencies" (SECs) to describe the set of competencies classified as interpersonal or intrapersonal in the National Research Council framework. In this article, we summarize evidence related to measurement of SECs.

The phrase "social and emotional competencies" and the related "social and emotional learning" (SEL) were popularized in large part by the widely used competency framework from the Collaborative for Academic, Social, and Emotional Learning (CASEL).[4] CASEL identified five categories of competencies—self-awareness, self-management, social awareness, relationship skills, and responsible decision-making. Within each of these categories are specific skills or mindsets such as "managing one's emotions" and "anticipating and evaluating the consequences of one's actions." This set of competencies is quite broad, and it is important to clarify what it does and does

not include. SECs comprise the skills, knowledge, and mindsets that people need to succeed and thrive. They are related to, but distinct from, mental health, school or classroom climate, and other conditions for learning.[5]

Interest in measuring SECs has increased in large part because of the growing evidence base[6] indicating the importance of these competencies for a variety of outcomes including academic achievement, workplace success, and civic participation. Changes in the employment landscape over the past decade have increasingly favored workers with strong social skills.[7] Further, advances in artificial intelligence have been accompanied by speculation regarding the skills young people will need to navigate changes to the job market, the media landscape, and the nature of relationships, with some arguing that SECs will be more important than ever.[8] Whatever the motivation, interest in quantifying SECs has grown.

Measurement of SECs can serve a variety of purposes. At the system (state [SEA] or local education agency [LEA]) level, SEC measurement can support monitoring of students' development of these competencies, send signals about high-priority outcomes, support family and community engagement around school performance, and inform district and school improvement efforts. In classrooms, afterschool programs, or other settings where young people develop SECs, measurement can inform decisions about curriculum and instruction and can empower students and educators to review progress and determine next steps. SEC measures also play a crucial role in research and evaluation of interventions.

Despite the demonstrated importance of SECs and the value of high-quality measurement for informing policy and practice related to SEC development, the availability of technically sound measures is uneven across competencies and across potential use cases. Research and development of SEC measures is widespread,[9] and emerging technologies such as AI offer promising avenues for assessment of hard-to-measure competencies.[10] But many of those efforts are nascent and have yet to be deployed extensively in real-world settings.

Key findings

Key finding #1: *Evidence of validity, reliability, and fairness should be gathered to support the use of SEC measures for a specific purpose.* These three foundational considerations are critical to evaluating the appropriateness of any use of a measurement tool in education. SEC measures pose specific challenges related to these considerations, including a susceptibility to gaming and a lack of evidence to support their use across diverse academic, social, and cultural contexts.

Key finding #2: *The best approach to assessing social and emotional competencies depends on the specific competency.* Approaches to measuring SECs fall into five broad categories: direct skills assessment, self-assessment of competencies, ratings by others, administrative data, and process/metadata (e.g., time spent on items). These approaches provide different information and are suitable for different competencies. For example, ratings by others can be used to assess competencies evident in observable behaviors. Self-assessments are especially valuable for assessing mindsets

or beliefs and are commonly used for large-scale monitoring but have potential biases, and scoring decisions can affect the validity and utility of scores.

Key finding #3: *Emerging technologies and methods can improve and personalize SEC measurement.* Emerging technological and methodological innovations show potential to transform how SEC measures are constructed and scored. For example, large language models (LLMs) have shown promise not only in writing quality items but also in identifying their psychometric properties, scoring item responses, and writing related reports. These benefits of LLMs can further be combined with existing or emerging methods to improve measurement. For instance, computer adaptive testing (CAT) paradigms match items to individuals in personalized ways that reduce administration time while improving reliability, but they rely on large item pools and costly large-scale item administrations. LLMs could help create larger item pools and avoid some piloting burden, putting CAT into play for SEC measures.

Key finding #4: *SEC definitions and measurement must be developmentally appropriate.* Research demonstrates that the nature of SECs changes over the developmental trajectory. Like other competencies, SECs become increasingly sophisticated as students progress from elementary to high school. Construct definitions and measurement methods need to align with students' developmental phases to the extent possible. Achieving this alignment is challenging, however, because of the lack of empirically based learning progressions for SECs. Additional research is needed to inform the creation of learning progressions that can be applied to SEC measurement.

Key finding #5: *Best practices are emerging for how to produce and use scores from SEC measures for research and policy analysis.* Research shows that approaches to scoring SEC measures can have large effects on results stemming from scores, especially when the scores are used to understand how children develop on these competencies and which programs to promote SECs prove effective. Thus, best practices are emerging for how to score SEC measures for such uses. That said, research also shows that scores—however they are produced—should not be used in isolation for decision-making about students, especially when the decisions are consequential.

Key finding #6: *Evidence to support the use of SEC measures for high-stakes purposes is limited.* Experts caution against the use of SEC measures in state accountability systems or for other high-stakes purposes. Almost no existing SEC measure has been validated for such uses. As SEAs and LEAs have sought to expand school performance measures to reflect the competencies students will need after graduation, many have adopted portraits of graduate or other frameworks that describe cross-cutting competencies similar to SECs, but for the most part, they are not yet measuring these competencies on a large scale; nor are they attaching high stakes for individuals or schools to these measures.

Evidence

Key finding #1: *Evidence of validity, reliability, and fairness should be gathered to support the use of SEC measures for a specific purpose.*

For any application of measurement, users are responsible for examining three foundational aspects of quality: validity, reliability, and fairness. The Standards for Educational and Psychological Testing, an authoritative set of professional guidelines, define validity as "the degree to which evidence and theory support the interpretations of test scores for proposed uses of tests."[11] A critical feature of this definition is the reference to tying evidence to a specific use. A measure developed and validated to inform classroom instruction or provide data for a research study, for example, should not be used in a high-stakes accountability system or for selection of individuals for admission to institutions or programs without validity evidence to support those uses. Put differently, validity is not a trait or property of a measure. New uses of a given measure—no matter how well known or how often cited the original validity study might be—necessitate additional evidence of validity to support that new intended use.

One particular threat to the validity of some SEC measures, such as those that rely on self-reports or educator ratings of behavior, is their susceptibility to gaming (i.e., deliberate manipulation of responses to achieve more favorable results) and to various biases (e.g., reference bias or social desirability bias). For instance, research shows that ratings of teachers or schools based on student surveys can be biased by classroom composition and noninstructional factors.[12] A related validity issue is the academic or content-related context in which SECs are measured. Students who have developed strong collaborative problem-solving skills in a science course might struggle to display those skills in an English course. Users of SEC measures should consider the benefits and risks associated with particular uses and should proceed cautiously if available evidence does not support those uses.

Reliability refers to "the degree to which test scores for a group of test takers are consistent over repeated applications of a measurement procedure and hence are inferred to be dependable and consistent for an individual test taker; the degree to which scores are free of random errors of measurement for a given group."[13] Because reliability is linked to measurement error, it is important to consider the potential sources of error relevant to a specific measure. For instance, a rubric that requires an adult to rate a child's skills or behavior is subject to measurement error stemming from rater effects: One rater might apply the rubric more stringently than another. (As an aside, this point was made to great effect in the Measures of Effective Teaching [MET] project, which showed that well under 50% of the variance in ratings of teacher effectiveness was at the level of the teacher, with much of the variance attributable to the rater.[14]) Evaluations of reliability for self-report survey-based SEC measures typically examines the internal consistency of the items—that is, the extent to which individual item scores correlate with scores on the other items. These reliability measures (e.g., Cronbach's alpha) capture information about measurement error due

to item sampling but do not provide evidence related to other sources of error such as occasion or context (for example, whether the student's SECs are being evaluated at home or in school).

Fairness refers to "the validity of test score interpretations for intended use(s) for individuals from all relevant subgroups."[15] Although an SEC measure might have evidence of validity based on previous administrations of the measure, that evidence might not be relevant to all populations and settings in which the measure is intended to be used. Students' approaches to applying competencies are often influenced by their social and cultural contexts. Students in societies that are more collectivist, for instance, might be likelier to refrain from expressing disagreement or to conform to cultural expectations when expressing emotions than students in more individualistic societies.[16] Relatively few SEC measures are supported by evidence of validity and reliability gathered across multiple social and cultural contexts, though one prominent exception is the Organisation for Economic Co-operation and Development's Survey on Social and Emotional Skills.[17]

Much like for academic test scores, valid uses of SEC scores for various purposes should generally involve combining scores with other data beyond those from SEC measures in decision-making over practice or policy. More specifically, they should be used in tandem with other data sources—such as classroom observations, academic performance, and school administrative data—for decision-making about students, teachers, or schools, particularly when the stakes are elevated. As noted above, SEC measures, especially self-report scales, contain measurement error, are sensitive to response styles, and can be influenced by factors unrelated to the underlying construct.[18] Moreover, psychometric studies of SEC scales caution that their reliability and validity vary across contexts and subgroups, meaning that classifications or decisions based on a single score may be unstable or misleading.[19] Consequently, best practice in educational measurement emphasizes multiple measures—triangulation of SEC data with behavioral indicators, academic metrics, and qualitative evidence—to support fair, accurate, and context-sensitive decision-making.[20]

Key finding #2: *The best approach to assessing social and emotional competencies depends on the specific competency.*

As mentioned briefly in the context of reliability, a variety of approaches to SEC measurement are available. For example, one could ask students to directly report on their SECs, try to have them demonstrate such skills directly, or have other people such as parents or teachers rate those skills. Each of these approaches has limitations. In this section, we describe various approaches that are most common, including tradeoffs associated with their use.

Self-assessment of competencies (self-report)

Self-report measures are perhaps the most common approach to measuring SECs. For example, the CORE districts in California administer self-report measures of SECs to more than 1 million students every year.[21] Self-report measures typically consist of brief, "fixed-form" survey items—often featuring Likert-type scales—in which

students evaluate their own thoughts, feelings, or behaviors on a consistent set of items. Because they draw directly on students' internal self-perceptions, they can efficiently capture information that no observer could easily infer. Duckworth and Yeager[22] emphasize that self-reports are inexpensive, scalable, and straightforward to administer, which makes them appealing for large-scale SEC data collection.

However, such measures have major limitations: Self-reports are vulnerable to social desirability pressures, faking, and reference group effects, where students compare themselves to different peer norms.[23] As an example of reference group effects, internationally, students compare themselves to peers within countries or even within schools in a way that undermines comparisons across countries. These biases can distort score meaning across classrooms or schools, especially under high-stakes conditions, reducing their usefulness for accountability or teacher evaluation.[24]

Ratings by others

An alternative to individuals reporting on themselves is other people with knowledge of the students reporting on them. For example, parents and teachers might report on a student's social awareness. Teacher report measures are common for SECs in education: Educators rate students' behaviors or personal qualities on the basis of sustained classroom observations, usually through structured rating scales. This approach leverages the fact that teachers witness students across tasks, days, and social situations, which allows them to aggregate impressions over time. Duckworth and Yeager[25] note that teacher reports can mitigate some shortcomings of self-reports by providing an external perspective less influenced by students' subjective standards.

At the same time, teacher ratings are prone to their own sources of bias, including halo effects, varying expectations across classrooms, and systematic differences in rating tendencies.[26] Because these contextual influences vary across settings, teacher report data should not be used for high-stakes comparisons of teachers or schools without significant caution. For example, research shows that teacher and parent ratings of children's SECs can differ substantively.[27] On one hand, such differences may be due not to rater bias but to how students behave at home versus in school (i.e., to important contextual differences). On the other hand, evidence indicates that rating differences relate to the race of the student and other background factors, suggesting bias could indeed be at play.[28]

Direct skills assessment

In contrast to self-report measures, direct assessments of SECs use structured tasks or behavioral simulations—such as delay-of-gratification tests or computerized impulse control tasks—to elicit observable behaviors linked to underlying competencies. These tasks aim to sidestep subjective judgment by measuring actual performance under standardized conditions. Such tasks help avoid many of the response biases associated with surveys and may offer more behaviorally grounded indicators of self-regulation or related skills.[29] Nevertheless, these assessments face serious practical and psychometric constraints: They are time intensive, costly to

administer at scale, and often noisy at the individual level, and they typically lack evidence that task responses generalize well to real-world contexts.[30]

Administrative data

Another source of data germane to SECs is school administrative data. (The approach of measuring SECs from administrative data has been considered for the workforce, as well.)[31] For example, attendance is relevant to how engaged a student is in school, with disengaged students likelier to be chronically absent.[32] In fact, given some of the challenges associated with self-report and direct skills assessment, many states' Every Student Succeeds Act (ESSA) plans use attendance as a nonacademic accountability metric.[33] However, such data are at best a proxy for the SEC of interest. Regarding the attendance example, while disengaged students are less likely to attend, there are myriad reasons a student might become chronically absent, including some beyond the student's direct control. Thus, such data should be used in tandem with other measures.

Process/metadata

Yet another option is to capture metadata relevant to SECs during other activities. For instance, research shows that, on computer-based achievement tests, the amount of time students spend on items—especially ones that are difficult for them—correlates highly with SECs such as academic motivation.[34] A range of similar findings has also accrued in the context of workforce preparedness.[35] A review of the literature found across multiple studies that such metadata predicted not only outcomes such as high school completion but even later earnings.[36] Researchers have further examined school contributions to these metadata and found large school-level effects.[37] However, as with administrative data, a major limitation of metadata is that such information is, at best, an imperfect proxy for the construct of interest and therefore should not be used in isolation. Nonetheless, metadata can provide an important piece of information to supplement and validate more traditional scores.

Key finding #3: *Emerging technologies and methods can improve and personalize SEC measurement.*

Emerging technological and methodological innovations show considerable potential to transform how SEC measures are constructed and scored. LLMs, for example, have shown promise to change many aspects of how SECs are created. Recent research demonstrates the utility of LLMs in the realm of item writing.[38] Although the study does not directly address SECs, it shows that the authors' AI framework can produce a large candidate pool of personality-scale items with strong linguistic acceptability, then uses neural models to estimate relationships among items and internal structure—thus creating items that perform similarly to their human-written equivalents. As an example of a concrete AI tool created to support item writing,[39] this approach was extended through AI-GENIE, an R statistical package that combines generative AI with network psychometrics to automatically generate and vet items for psychological constructs. These advances suggest that LLMs can likely meaningfully reduce the human burden of item generation, a longstanding bottleneck in SEC measurement,

including in cases such as adaptive assessment where large, diverse item banks are essential.

In addition to item writing, LLMs and AI can predict item performance prior to large-scale piloting. For example, researchers[40] have used neural networks to forecast which generated items would cohere into scales with high reliability. Other research has used LLMs along with network modeling to evaluate which AI-generated items were semantically and statistically coherent with one another—i.e., form a consistent set of items measuring the same idea or construct (such as empathy or self-control).[41] These applications of LLMs suggest that much of the costly and time-consuming human piloting required to generate new measures could be reduced if not eliminated. Furthermore, researchers have applied machine learning algorithms to interpersonal assessments and found that these algorithms could largely reproduce human assessor ratings, providing initial evidence of automated scoring viability.[42] These findings imply that AI can both help determine item psychometric properties and score those items, thereby accelerating scale development and freeing resources for other measurement tasks.

These LLM capabilities solve some of the largest challenges in making SEC measures more personalized through adaptation. Adaptive measures produce a real-time estimate of the individual's score on an outcome such as an SEC and then match an item to that score for each question. Such personalization has several benefits, including requiring many fewer items without sacrificing reliability. However, adaptive frameworks—often referred to as computer-adaptive tests (CATs)—rely on very large pools of items with known properties. Creating these item pools requires extensive pilot testing, which is time- and resource-intensive. If LLMs can help write the items and identify their psychometric properties, then SEC-based CATs can be created straightforwardly. There is also software available with a built-in survey response interface such that building CATs is far easier than in the past.[43] Meanwhile, emerging research shows that CAT methods can reduce administration time and maintain precision even in surveys in which most respondents agree with many items—a challenge typical of SEC self-reports.[44]

In addition to CAT applications, LLMs have the potential to enable other personalization approaches such as tailoring of content to students' motivational levels, development of items relevant to diverse social and cultural contexts, and personalization of feedback as part of a formative assessment process.[45] However, most of these applications have not yet been widely tested in diverse student populations or contexts.

Key finding #4: *SEC definitions and measurement must be developmentally appropriate.*

Developmental science demonstrates that SECs evolve unevenly and nonlinearly from early childhood through adolescence. To produce valid, actionable information for policymakers and educators, SEC measurement must be aligned with what children and adolescents are actually capable of at different stages of development. Decades of work in developmental psychology shows that competencies such as recognition of emotions, responsible decision-making, self-regulation, and social problem-solving

emerge at different ages and follow distinct developmental paths.[46] For instance, young children begin by identifying simple emotions before progressing toward more complex skills, such as interpreting mixed emotions or understanding that people can mask feelings.[47] Maturation in one domain does not automatically translate into growth in another, so it is important to understand the developmental trajectories of the specific competencies being measured.

Because of these developmental patterns, the method of measurement must match the student's age and developmental capacities to the extent possible because the meaning of a competency, and the way it is expressed, can change over the lifespan.[48] Developmental considerations point to the importance of matching the method of measurement to the grade level. In preschool and early elementary grades, children lack the reading ability and self-reflective capacity needed for accurate self-reporting. For this developmental stage, observational rubrics and teacher or parent rating scales are likely to provide more accurate evidence of naturally occurring behaviors, including cooperation and basic emotion regulation, than self-report.[49]

Upper-elementary students can begin responding to carefully designed self-report questionnaires provided the reading level is appropriate and the items use concrete, recent, and context-specific wording ("In the last week, I..."). Studies provide some evidence of validity for scenario-based or performance tasks that target skills such as conflict resolution or emotion recognition among children in this age range.[50] By middle and high school, adolescents possess stronger metacognitive skills and a stabler sense of identity, which make self-report more feasible for constructs such as belonging, growth mindset, and self-efficacy. However, it is important to keep in mind, as noted above, that self-reports are sensitive to social desirability and reference biases. Further, such biases may change over time as students age.[51] And of course, within any age or grade level, students' capabilities for responding to particular types of assessment are likely to vary.

When researchers examine SECs over time, such as when they study the long-term effects of an intervention on growth in SECs, it is important to consider measurement invariance. In other words, the measures of interest must function the same way across different ages and developmental stages. For example, if the reliability of a given item (quantified by means of item weights such as factor loadings and discrimination parameters) differs with respondent age, then failing to account for such differences can bias interpretation of scores. In the absence of evidence of invariance, any observed changes may reflect shifts in respondents' interpretation of a measure or its meaning to them rather than true growth.[52]

One challenge facing researchers and policymakers interested in measuring SECs longitudinally is the lack of evidence-based learning progressions for these competencies. A learning progression describes a pathway of increasing sophistication in knowledge, skills, and behaviors within a domain, outlining what learners typically know and can do at successive stages on the way to a more advanced target.[53] Learning progressions help connect developmental theory, assessment design, instructional planning, and formative uses of data by showing what prior capacities learners must build before more advanced ones emerge. Learning

progressions are not tied explicitly to age or grade level; an 8-year-old and a 12-year-old might demonstrate skills associated with the same point along a learning progression. While extensive work has been conducted to develop learning progressions in core academic subjects,[54] there are almost no empirically validated learning progressions for SECs.

Key finding #5: *Best practices are emerging for how to produce and use scores from SEC measures for research and policy analysis.*

Research shows that approaches to scoring SEC measures—especially when they take the form of brief survey instruments—can have consequences for the understanding of broader research and policy questions related to those competencies.[55] For example, scores produced by simple adding-up of the item responses (so-called sum scores) tend to understate not only growth trajectories and the effects of interventions but also variability in those trajectories and effects across students.[56] Emerging best practices indicate that the statistical models used to score such measures might ideally match the nature of the study design.[57] Results from a pre/post randomized control trial, for instance, might use a statistical model for scoring that matches the multi-group (control and treatment) and multi-timepoint (pre/post) nature of the data. Such approaches are likelier to match true developmental patterns and intervention effects on those patterns. A tutorial on how to approach these scoring decisions is available from Soland and colleagues.[58]

However, as discussed in the section on validity, research also suggests that even when SEC measures are scored according to best practices, the scores should often be paired with other data for the drawing of inferences or policy analysis.[59] One reason to pair scores with other data relates to practical significance. For example, if a growth mindset intervention or program results in an SEC score increase of a .20-standard-deviation unit, how does one conclude whether such a change is meaningful and therefore worth the investment? Fortunately, tools exist to help evaluators understand whether impacts are practically meaningful. On one hand, policymakers, researchers, and educators can turn to meta-analyses, which summarize effects of interventions on SECs across studies to better contextualize the magnitude of such effects. One recent meta-analysis[60] shows that, compared to students in the control conditions, students who participated in schoolwide interventions to boost SECs saw significant gains in skills, attitudes, behaviors, school climate and safety, peer relationships, school functioning, and academic achievement. Such results can be used as a point of comparison for specific results from other contexts. On the other hand, SEC scores can be compared to empirical benchmarks for related outcomes.[61] These benchmarks quantify what "typical" gains or growth look like by student age or grade level and then allow stakeholders to compare impacts from their own programs or policies to what is typical. In the realm of SEC measures, two recent studies[62] try to produce such benchmarks for gains and growth over time, respectively, in SEC outcomes and offer important caveats about when such benchmarks are and are not appropriate for comparisons.

Key finding #6: *Evidence to support the use of SEC measures for high-stakes purposes is limited.*

Most researchers and measurement experts who have studied and written about SEC measures have urged users to avoid attaching high stakes to the use of such measures.[63] These concerns are likely one factor contributing to the absence of SEC measures in states' accountability systems under the ESSA.[64] The primary reason for arguing against high-stakes uses is the lack of evidence of validity to support such uses. Despite the growing interest in measuring SECs in educational settings, most measures have not been validated for high-stakes uses such as admissions, employment, and accountability. One factor that makes SEC measures particularly unsuitable for high-stakes uses is their reliance on self-report surveys, which are easily subject to score inflation.[65]

California's CORE Districts provide an example of an effort to use SEC measures in an accountability system without attaching high stakes to their use.[66] Under a waiver from federal accountability requirements in 2013, the CORE districts (comprising eight California districts at the time) adopted an accountability system that incorporated measures of SECs and school climate in addition to academic outcomes. Data from an SEC survey were used for monitoring and capacity-building purposes, but not for high-stakes accountability.[67] The SEC data proved useful for guiding school and district improvement. For instance, one study finds that school-level changes in scores on the SEC surveys were associated with changes in student achievement and attendance.[68]

The lack of SEC measures suitable for accountability purposes poses challenges for educators and policymakers who want to ensure that schools are preparing students for postsecondary success in the context of economic, technological, and societal changes. As SEAs and LEAs work to identify and cultivate critical competencies among their students, they have explored new ways of defining and measuring student outcomes. For example, many states and districts have adopted portraits of a graduate that describe the competencies all graduates are expected to demonstrate and that are intended to guide instruction and curriculum decisions. For the most part, however, assessments that can measure these competencies at large scale, across diverse populations and community contexts, and in a feasible and cost-effective manner do not exist. Instead, assessment of SECs in these systems typically relies on locally developed rubrics that educators or students complete.[69] These rubrics, especially when tied to high-quality, curriculum-aligned tasks, can provide useful information for guiding instruction. But they are not well suited for large-scale data collection that requires comparability.

Endnotes and references

[1] America Succeeds. 2025. Durable by Design: An Update on the High Demand for Durable Skills; Terrones, Felipe, and Matthew Reeves. 2025. Creating New and Better Ways to Measure the "Soft Skills" Students Need to Succeed. Urban Institute.
[2] National Research Council. 2012. Education for Life and Work: Developing Transferable Knowledge and Skills in the 21st Century. National Academies Press.
[3] Harvard EASEL Lab. Navigate the Complex Field of Social and Emotional Learning.
[4] Collaborative for Academic, Social, and Emotional Learning. What Is the CASEL Framework?; Edutopia. 2011. Social and Emotional Learning: A Short History, October 6.
[5] McKown, Clark, and Maria Kharitonova. 2025. Assessing SEL: Perspectives from the Field. In Handbook of Social and Emotional Learning, 2nd ed. Edited by Joseph A. Durlak et al. Guilford Press.
[6] Cipriano, Christina, et al., 2023. The State of Evidence for Social and Emotional Learning: A Contemporary Meta-analysis of Universal School-based SEL Interventions. Child Development 94(5): 1181–204.
[7] Deming, David J. 2017. The Growing Importance of Social Skills in the Labor Market. The Quarterly Journal of Economics 132(4): 1593–640.
[8] Prothero, Arianna. 2023. Artificial Intelligence and Social-Emotional Learning Are on a Collision Course. Education Week, November 13.
[9] Burrus, Jeremy, et al. 2023. Assessing Competencies for Social and Emotional Learning: Conceptualization, Development, and Applications. Routledge; Jones, Stephanie M., et al. 2022. Measuring Noncognitive Skills in School Settings. Guilford Press; Terrones and Reeves, 2025.
[10] Organisation for Economic Co-operation and Development (OECD). 2024. Nurturing Social and Emotional Learning Across the Globe: Findings from the OECD Survey on Social and Emotional Skills 2023. OECD Publishing.
[11] American Educational Research Association (AERA) et al. 2014. Standards for Educational and Psychological Testing, p. 11. AERA.
[12] Ferguson, Ronald F., and Charlotte Danielson. 2015. How Framework for Teaching and Tripod 7Cs Evidence Distinguish Key Components of Effective Teaching. In Designing Teacher Evaluation Systems, 1st ed. Edited by Thomas J. Kane et al. Wiley; Kuhfeld, Megan, et al. 2023. Investigating Differences in How Parents and Teachers Rate Students' Self-Control. Psychological Assessment 35(1): 23; Kraft, Matthew A. 2020. Interpreting Effect Sizes of Education Interventions. Educational Researcher 49(4): 241–53; Edwards, Kelly, and James Soland, 2024. Improving the Precision of Classroom Observation Scores Using a Multi-Rater and Multi-Timepoint Item Response Theory Model. Educational Assessment 29(2): 103–23; Ho, Andrew D., and Thomas J. Kane, 2013. The Reliability of Classroom Observations by School Personnel. Research Paper. MET Project. Bill & Melinda Gates Foundation, ERIC.
[13] AERA et al., 2014, pp. 222–223.
[14] Kane, Thomas J., and Douglas O. Staiger. 2012. Gathering Feedback for Teaching: Combining High-Quality Observations with Student Surveys and Achievement Gains. Bill & Melinda Gates Foundation.
[15] AERA, et al., 2014, p. 219.
[16] Vishkin, Anat, Shinobu Kitayama, M. Kyle Berg, Ed Diener, Dana Gross-Manos, Asher Ben-Arieh, and Maya Tamir. 2023. Adherence to Emotion Norms Is Greater in Individualist Cultures Than in Collectivist Cultures. Journal of Personality and Social Psychology 124(6): 1256–1276.
[17] OECD, 2024.
[18] Duckworth, Angela L., and David Scott Yeager. 2015. Measurement Matters: Assessing Personal Qualities Other than Cognitive Ability for Educational Purposes. Educational Researcher 44(4): 237–51; Krosnick, Jon A. 1999. Survey Research. Annual Review of Psychology 50(1): 537–67; Soland, James, and Megan Kuhfeld. 2020. Do Response Styles Affect Estimates of Growth on Social-Emotional Constructs? Evidence from Four Years of Longitudinal Survey Scores. Multivariate Behavioral Research, 1–21.

[19] AERA et al., 2014.
[20] AERA et al., 2014.
[21] Gehlbach, Hunter, and Heather J. Hough. 2018. Measuring Social Emotional Learning through Student Surveys in the CORE Districts: A Pragmatic Approach to Validity and Reliability. Policy Analysis for California Education (PACE), ERIC.
[22] Duckworth and Yeager, 2015, p. 237–251.
[23] Heine, Steven J., et al. 2002. What's Wrong with Cross-Cultural Comparisons of Subjective Likert Scales? The Reference-Group Effect. Journal of Personality and Social Psychology 82(6): 903.
[24] Heine et al., 2002.
[25] Duckworth and Yeager, 2015.
[26] Ready, Douglas D., and David L. Wright. 2011. Accuracy and Inaccuracy in Teachers' Perceptions of Young Children's Cognitive Abilities: The Role of Child Background and Classroom Context. American Educational Research Journal 48(2): 335–60.
[27] Kuhfeld et al., 2023.
[28] Kuhfeld et al., 2023.
[29] Duckworth and Yeager, 2015; Soland, Jim, et al. 2013. Measuring 21st Century Competencies: Guidance for Educators, RAND Corporation.
[30] Eisenberg, Nancy, et al. 2010. Empathy-Related Responding: Associations with Prosocial Behavior, Aggression, and Intergroup Relations: Empathy-Related Responding. Social Issues and Policy Review 4(1): 143–80.
[31] Kyllonen, Patrick C., et al. 2025. SEL Competencies in the Workplace, in Handbook of Social and Emotional Learning, 2nd ed. Edited by Joseph A. Durlak et al. Guilford Press.
[32] Soland, James, et al. 2019. Are Test and Academic Disengagement Related? Implications for Measurement and Practice. Educational Assessment 24(2): 1–16; Gottfried, Michael A. 2014. Chronic Absenteeism and Its Effects on Students' Academic and Socioemotional Outcomes. Journal of Education for Students Placed at Risk 19(2): 53–75; Archambault, Isabelle, et al. 2009. Student Engagement and Its Relationship with Early High School Dropout. Journal of Adolescence 32(3): 651–70.
[33] Bauer, Lauren, et al. 2018. Reducing Chronic Absenteeism under the Every Student Succeeds Act. Brookings Institution, p. 1–31.
[34] Soland, James. 2019. Can Item Response Times Provide Insight Into Students' Motivation and Self-Efficacy in Math? An Initial Application of Test Metadata to Understand Students' Social–Emotional Needs. Educational Measurement: Issues and Practice 38(3): 86–96.
[35] Kyllonen et al., 2025; Kyllonen, Patrick C. 2015. Designing Tests to Measure Personal Attributes and Noncognitive Skills. In Handbook of Test Development. Routledge.
[36] Soland, James, et al. 2019. Identifying Naturally Occurring Direct Assessments of Social-Emotional Competencies: The Promise and Limitations of Survey and Assessment Disengagement Metadata. Educational Researcher 48(7): 466–78.
[37] Jensen, Nate, et al. 2018. The Influence of Rapidly Guessed Item Responses on Teacher Value-Added Estimates: Implications for Policy and Practice. Educational Evaluation and Policy Analysis 20(1): 90–98; Baird, Matthew, et al. 2025. Beyond Test Scores: The Impact of Test Persistence Value-Added on Student Achievement. Education Economics 33(6): 885–903.
[38] Hernandez, Ivan, and Weiwen Nie. 2023. The AI-IP: Minimizing the Guesswork of Personality Scale Item Development through Artificial Intelligence. Personnel Psychology 76(4): 1011–35.
[39] Russell-Lasalandra, Lara, et al., n.d. Generative Psychometrics via AI-GENIE: Automatic Item Generation with Network-Integrated Evaluation. Open Science Foundation. Accessed November 30, 2025.
[40] Hernandez and Nie, 2023.
[41] Russell-Lasalandra et al., n.d.
[42] Hickman, Louis, et al. 2023. Automatic Scoring of Speeded Interpersonal Assessment Center Exercises via Machine Learning: Initial Psychometric Evidence and Practical Guidelines. International Journal of Selection and Assessment 31(2): 225–39.
[43] Chalmers, R. Philip. 2016. Generating Adaptive and Non-Adaptive Test Interfaces for Multidimensional Item Response Theory Applications. Journal of Statistical Software 71: 1–38.
[44] Soland, James. 2025. Exploring the Feasibility and Usefulness of Creating Computer Adaptive Surveys When Respondents Agree to Most Items. Educational Assessment, October 6: 1–19.

[45] Arslan, Burcu, Brent Lehman, Charles Tenison, J. R. Sparks, A. A. López, Li Gu, and Diego Zapata-Rivera. 2024. Opportunities and Challenges of Using Generative AI to Personalize Educational Assessment. Frontiers in Artificial Intelligence 7: 1460651.
[46] Barbarin, Oscar A., and Barbara Hanna Wasik (Eds.) 2009. Handbook of Child Development and Early Education: Research to Practice. Guilford.
[47] Denham, Susanne A. 2007. Dealing with Feelings: How Children Negotiate the Worlds of Emotions and Social Relationships. Cognition, Brain, Behavior 11(1): 1–48.
[48] Denham, Susanne A. 2018. Keeping SEL Developmental: The Importance of a Developmental Lens for Fostering and Assessing SEL Competencies. Collaborative for Academic, Social, and Emotional Learning (CASEL).
[49] Assessment Work Group. 2019. Student Social and Emotional Competence Assessment: The Current State of the Field and a Vision for Its Future. CASEL.
[50] Russo-Ponsaran, Nicole M., et al. 2022. Novel Technology-Based SEL Assessments. In Assessing Competencies for Social and Emotional Learning: Conceptualization, Development, and Applications. Edited by Jeremy Burrus et al. Routledge.
[51] Soland and Kuhfeld, 2020.
[52] Lee, Juyeon, et al. 2023. Measuring the Development of Social-Emotional Competence Using Behavioral Rating Scales in the Context of School-Based Social and Emotional Learning. Social and Emotional Learning: Research, Practice, and Policy 2.
[53] Heritage, Margaret. 2008. Learning Progressions: Supporting Instruction and Formative Assessment. Paper Prepared for the Formative Assessment for Teachers and Students (FAST) SCASS (Council of Chief State School Officers).
[54] Daro, Phil, et al. 2011. Learning Trajectories in Mathematics: A Foundation for Standards, Curriculum, Assessment, and Instruction. CPRE Research Report RR-68. Consortium for Policy Research in Education.
[55] McNeish, Daniel, and Melissa Gordon Wolf. 2020. Thinking Twice about Sum Scores. Behavior Research Methods, p. 1–19.
[56] Kuhfeld, Megan, and James Soland. 2022. Avoiding Bias from Sum Scores in Growth Estimates: An Examination of IRT-Based Approaches to Scoring Longitudinal Survey Responses. Psychological Methods 27(2): 234; Soland, James, et al. 2024. How Survey Scoring Decisions Can Influence Your Study's Results: A Trip through the IRT Looking Glass. Psychological Methods 29(5): 1003; Soland and Kuhfeld, 2020; Edwards, Kelly D., and James Soland. 2024. How Scoring Approaches Impact Estimates of Growth in the Presence of Survey Item Ceiling Effects. Applied Psychological Measurement 48(3): 147–64; Soland, James. 2022a. Evidence That Selecting an Appropriate Item Response Theory–Based Approach to Scoring Surveys Can Help Avoid Biased Treatment Effect Estimates. Educational and Psychological Measurement 82(2): 376–403.
[57] Gorter, Rosalie, et al. 2016. Measurement Model Choice Influenced Randomized Controlled Trial Results. Journal of Clinical Epidemiology 79: 140–49; Bauer, D., and P. Curran. 2015. The Discrepancy Between Measurement and Modeling in Longitudinal Data Analysis. Advances in Multilevel Modeling for Educational Research: Addressing Practical Issues Found in Real-World Applications, p. 3–38; Kuhfeld and Soland, 2022; Soland et al., 2024; Soland, 2022a.
[58] Soland et al., 2024.
[59] Kraft, 2020; Kraft, Matthew A. 2023. The Effect-Size Benchmark That Matters Most: Education Interventions Often Fail. Educational Researcher 52(3): 183–87.
[60] Cipriano et al., 2023.
[61] Bloom, Howard S., et al. 2008. Performance Trajectories and Performance Gaps as Achievement Effect-Size Benchmarks for Educational Interventions. Journal of Research on Educational Effectiveness 1(4): 289–328; Hill, Carolyn J., et al. 2008. Empirical Benchmarks for Interpreting Effect Sizes in Research. Child Development Perspectives 2(3): 172–77; Lipsey, Mark W., et al. 2012. Translating the Statistical Representation of the Effects of Education Interventions Into More Readily Interpretable Forms. Institute of Education Sciences, US Department of Education; Kraft, 2020.
[62] Soland, James, et al. 2022b. Empirical Benchmarks for Changes in Social and Emotional Skills over Time. Child Development 93(4): 1129–44; Rimm-Kaufman, Sara E., et al. 2024. Social and Emotional Competency Development from Fourth to 12th Grade: Relations to Parental Education and Gender. American Psychologist. American Psychological Association.
[63] Duckworth and Yeager, 2015; Melnick, Hannah, et al. 2017. Encouraging Social and Emotional Learning in the Context of New Accountability. Learning Policy Institute;

Engineering National Academies of Sciences and Medicine. 2019. Monitoring Educational Equity. National Academies Press.
[64] Blad, Evie. 2017. No State Will Measure Social-Emotional Learning under ESSA. Will That Slow Its Momentum? Education Week.
[65] Duckworth and Yeager, 2015; Melnick et al., 2017.
[66] Gehlbach and Hough, 2018.
[67] West, Martin R., et al. 2018. Trends in Student Social-Emotional Learning: Evidence from the CORE Districts. Working Paper. PACE.
[68] Kanopka, Kathleen, et al. 2024. Are Changes in Reported Social-Emotional Skills Just Noise? The Predictive Power of Longitudinal Differences in Self-Reports. AERA Open 10.
[69] Dunbar, Kyle, et al. 2025. Portraits in Progress: Emerging Future-Ready Skills Assessment Strategies. Digital Promise and Pearson, 2025.