METHODOLOGICAL ARTICLE

Web-based assessment of social and emotional competence in the late elementary grades

Clark McKown[1,2] | **Nicole Russo-Ponsaran**[1] | **Ashley Karls**[1]

[1]Department of Psychiatry and Behavioral Sciences, Rush University Medical Center, Chicago, Illinois, USA

[2]xSEL Labs, Evanston, Illinois, USA

Correspondence
Clark McKown, Department of Psychiatry & Behavioral Sciences, Rush University Medical Center, RNBC 4711 Golf Road, Suite 1100, Skokie IL 60076, USA.
Email: Clark_A_McKown@rush.edu

Abstract

This paper presents evidence of the score reliability, factor structure, criterion-related validity, and measurement equivalence of a web-based assessment of several important social and emotional competencies for children in fourth through sixth grades. The assessment, SELweb LE (Late Elementary), is designed to measure children's understanding of other's thoughts and feelings, their ability to solve social problems, and their ability to engage in self-control. SELweb LE satisfies a need for a direct assessment that measures theoretically relevant social and emotional competencies specifically for youth in the late elementary grades. The competencies measured are associated with youth outcomes, are reflected in state education standards, and are the targets of instruction in widely used social and emotional learning (SEL) curricula. Analyses support SELweb LE's factor structure, score reliability, measurement equivalence, and convergent, discriminant, and criterion-related validity. More broadly, evidence supports the use of SELweb LE to characterize student competencies.

KEYWORDS
assessment, late elementary, middle childhood, social and emotional learning

1 | INTRODUCTION

The ability to understand and effectively interact with others includes competencies such as self-awareness, self-management, social awareness, relationship skills, and responsible decision-making (https://casel.org/fundamentals-of-sel/what-is-the-casel-framework/). These social and emotional (SE) competencies are critical for academic, social, and life success (DiPerna & Elliott, 2002; Payton et al., 2008) and they can be taught (Durlak et al., 2011). Recognizing

that SE competencies are meaningful and malleable, a growing number of states have adopted standards describing the SE competencies children should know and be able to demonstrate (Dusenbury et al., 2018). In addition, many effective programs are designed to nurture the SE competencies described in state standards (https://pg.casel.org/). In this context, a growing number of school districts have adopted some form of social and emotional learning (SEL) initiative. Two 2019 surveys found that more than half of American educators report engaging in SE instruction (Atwell & Bridgeland, 2019; Hamilton et al., 2019).

In this context, educators report wanting to use SE competence assessments to guide SE instruction and measure student growth, but are not sure how (Atwell & Bridgeland, 2019; Hamilton et al., 2019). Furthermore, researchers have called for investigation into new and useful ways to assess SE competencies (Duckworth & Yeager, 2015). To address the need for new assessments that address educators' needs, we created SELweb EE (Early Elementary) a SE competence assessment for kindergarten to third grade designed to support educational practice (McKown et al., 2016, 2021). We then aimed to extend the age range for which direct assessment is available and created SELweb LE (Late Elementary) for assessment of fourth to sixth graders. This paper describes the results of three studies examining SELweb LE's psychometric properties.

1.1 | The context for SELweb LE

Two commonly used methods of assessing SE competence in school include teacher rating scales and child self-report scales. With teacher rating scales, teachers rate how often a student engages in a range of behaviors. As such, rating scales are well-suited to assess observable SE competencies. With self-report, children rate the self-relevance of statements reflecting their own skills or dispositions, the frequency of their own behavior, or the ease of certain SE competencies. Self-report may be well-suited to assessing beliefs and attitudes, such as growth mindset (Dweck et al., 1995). Other competencies, such as children's ability to infer others' emotions and perspectives and solve social problems are thinking skills and may be less well-measured by surveys or rating scales. In the case of rating scales, observers may have difficulty seeing these competencies and thus may have difficulty accurately rating them. In the case of surveys, children may be unable or unwilling to accurately report their own skill level (Bernard et al., 1984; Crowne & Marlowe, 1960; Fan et al., 2006; Shrauger & Osberg, 1981).

In contrast to self-report and rating scales, which are indirect measures because they measure a respondent's *perceptions* of skill level, *direct assessments* measures skills by having children solve challenging problems in which they demonstrate their SE competence (Assessment Work Group, 2019). For example, one component of CASEL's "social awareness" involves understanding others' emotions. To measure this skill, on a self-report survey, respondents might rate on a Likert scale how frequently they can tell what another person is feeling from their facial expressions. With a teacher rating scale, teachers might rate how often a child understands what others are feeling. Both methods capture the respondent's *perceptions* of a child's skill level. With direct assessment, children might look at photographs of faces and label each person's emotions. For each item, there are correct and incorrect responses and in their responses to the items, children are thereby demonstrating their level of competence. A key motivation for developing SELweb LE is our belief that direct assessments can provide insights into children's SE competence that are otherwise difficult to ascertain.

Existing direct assessments engender strengths and limitations. Some assess specific competencies, including the extent to which children can understand others' feelings (Denham, 1986; Innovation Research & Training, Inc, 2008; Korkman et al., 2007; Nowicki & Duke, 1994; Pons et al., 2004), can infer others' perspectives (Korkman et al., 2007), can solve social problems (SIP-AP; Kupersmidt et al., 2011; VESIP; Russo-Ponsaran et al., 2021), or engage in self-control (Bitsakou et al., 2006; Duckworth & Kern, 2011; Kuntsi et al., 2001; Willoughby et al., 2011). Each of these existing assessments has merits. However, each is designed to measure a specific SE competence, omitting a range of other important competencies.

Two direct assessments that measure multiple SE competencies are the Mayer-Salovey-Caruso Emotional Intelligence Test—Youth Version (MSCEIT-YV; Mayer et al., 2005) and SELweb EE (McKown et al., 2016). The MSCEIT-YV is based on the theory of emotional intelligence (Salovey & Mayer, 1990) and assesses competencies such as emotion understanding and emotion management in youth the ages of 10–17 years old (Rivers et al., 2012). The MSCEIT-YV includes subtests that are designed to measure several emotion-specific competencies corresponding to the theory of emotional intelligence proposed by Salovey and Mayer (1990).

SELweb EE was developed to assess a range of SE competencies that are reflected in the CASEL model and state SEL standards, and that are targets of instruction in evidence-based SEL programs. Like the MSCEIT-YV, SELweb EE focuses on emotions and includes subtests designed to assess emotion recognition (part of CASEL's social awareness) and self-control (part of CASEL's self-management). In addition, SELweb EE includes subtests measuring social perspective-taking (SPT) (part of CASEL's social awareness) and social problem-solving (SPS) (part of CASEL's relationship skills and responsible decision-making). SELweb EE has been validated for kindergarten through third grades and uses age-appropriate graphics and voice-over narration (McKown et al., 2016, 2021). As evidence of its usefulness in authentic education settings, SELweb EE is used for characterizing competencies and measuring student progress in hundreds of schools nationally.

What is missing is a direct assessment for the late elementary grades that reflects SELweb EE's design features. To that end, SELweb LE was designed for the late elementary grades—fourth to sixth grades—to be developmentally appropriate and to assess SE competencies that are aligned to the CASEL model, are theoretically grounded, are associated with youth outcomes, are incorporated in state standards, and are the targets of instruction in evidence-based SEL programs.

1.2 | SELweb LE: Rationale and overview

Many models of SE competence have been proposed (Berg et al., 2017). Among the alternatives, the CASEL model is widely cited and is the basis for many state standards. The CASEL model highlights the importance of five broad competencies. Within each of these broad competencies, however, there are a multitude of sub-competencies and therefore ways those competencies could be operationalized. Because SE competence includes such an expansive range of skills (https://casel.org/core-competencies/), it would be difficult to assess every possible SE competence without creating a prohibitively lengthy assessment. In determining what SELweb LE should specifically measure, we opted to measure competencies that are: (a) reflected in the CASEL model, (b) strongly associated with important outcomes, and (c) the targets of instruction in evidence-based SEL programs. As a result, SELweb LE measures a subset of CASEL competencies. In addition, to provide continuity across grades, we elected to assess competencies assessed by SELweb EE. We provide the rationale for SELweb LE's construct coverage here, and detailed descriptions of the assessment subtests in Table 1 and the Study 1 methods section.

SELweb LE's measures two components of CASEL's "social awareness"—emotion recognition and SPT. The literature supports the conclusion that these two elements of social awareness are consequential. Nowicki and Duke (1994) reviewed 14 studies and found that among elementary-aged children, understanding others' basic emotions (joy, sadness, anger, and fear) was associated with reading and math achievement, peer acceptance, self-esteem, and locus of control. In the late elementary grades, children begin to understand that people can feel mixed emotions and complex social emotions such as pride, embarrassment, and guilt (Pons et al., 2004). For SELweb LE, therefore, we designed a vignette-based task to assess children's complex emotion recognition (CER) skill. In this task, children are presented with 12 situations in which a character might feel mixed or complex emotions and asked to select one or more emotions that the person would feel. Answer choices were Happy, Sad, Angry, Nervous, Proud, Embarrassed, and Guilty.

Furthermore, in the late elementary grades, understanding others' perspectives—another element of CASEL's social awareness—is associated with behavior, friendship quality, and peer acceptance (Diazgranados et al., 2016; Oberle, 2018; Smith & Rose, 2011). Accordingly, we designed a SPT subtest to assess children's ability to infer others'

TABLE 1 eScription of SELweb modules, questions, and item scoring

Module	CASEL competence	Item description	# Items	Question and response options	Item score	Possible range
Complex emotion recognition	Social awareness	Respondents listen to illustrated, narrated vignettes and are asked how the character feels.	12	How does the child feel? happy, sad, angry, nervous, proud, embarrassed, or guilty	1: Correct emotion 0: Incorrect emotion	0–24
Social Perspective Taking	Social awareness	Respondents listen to illustrated, narrated vignettes and answer questions.	21	Seven questions for each of three vignettes about what characters know about the situation and each other. Responses options are illustrated, narrated forced choice.	1: Correct mental state inference 0: Incorrect answer	0–21
Social problem solving	Relationship skills responsible decision-making	Respondents hear illustrated, narrated vignettes involving either ambiguous provocation, peer entry, or differences of opinion.	6	**Goal Preference** What do you want to happen? Response options are narrated forced choice that include retribution, domination, prosocial and avoidance goals.	2: Resolving the problem 1: Avoid the problem 0: Retribution or domination	0–12
			6	**Solution Preference** What will you do? Response options are narrated forced choice that include positive-assertive, passive, and aggressive solutions.	2: Positive-assertive 1: Passive 0: Aggressive	0–12
			6	**Consequence Prediction** If you [chosen action], what will happen? Response options include narrated forced choice positive or negative outcomes.	1: Accurate and competent 0: Inaccurate and/or incompetent	0–12
Emotion regulation knowledge	Self-management	Respondents hear illustrated narrated vignettes and are asked what the characters could do to feel better	12	What can the person do to feel better? Response options include illustrated, narrated forced choice, with four possible responses.	1: Correct answer 0: Incorrect answer	0–12
Self-reported self-control	Self-management	Children rate statements about their self-control skills.	7	Rate How True on a five-point Likert scale	1-5 points (1 = not at all true; 5 = extremely true)	7–35

intentions and perspectives. This subtest draws on the literature on children's theory of mind (Devine & Hughes, 2013; Happé, 1994) and SPT (Selman, 1980), which suggest that by the late elementary grades, children can infer what others think, including others' thoughts about the thoughts of a third party. To assess SPT, we designed a subtest in which children are presented with stories and asked questions about characters' words, actions, and intentions in each story. We hypothesized that the CER and SPT subtests together reflect a single construct we call "understanding others."

An additional SELweb LE subtest was designed to measure elements of the CASEL competencies "relationship skills" and "responsible decision-making." This SPS subtest draws from the literature on social information processing (e.g., Crick & Dodge, 1994). SPS is associated with behavioral and academic functioning from preschool through adolescence (Denham et al., 2012; Dubow et al., 1991; Nelson & Crick, 1999; Rotheram, 1987). Consistent with these literatures, we conceptualize SPS as a multi-component process in which, when confronting social situations, children are guided by social goals, generate, and select alternative solutions, and evaluate solutions in part by predicting their consequences. To assess SPS, we designed a subtest in which children are presented with hypothetical challenging social situations, asked to imagine they were in the situation, and asked about their goals (e.g., "Pick the statement that most closely matches what you want"), preferred actions (e.g., "What will you do?"), and the predicted consequences of those actions (e.g., "If you [do that], what will happen?"). Each question had a selection of answer options to choose from.

Two additional subtests were designed to assess components of CASEL's "self-management." In children the age of 9–18 years, self-control is positively associated with adaptive behavior and reading achievement, and negatively associated with emotional and behavioral challenges and disciplinary referrals (Peters et al., 2009; Rivers et al., 2012). One SELweb LE subtest is designed to assess children's knowledge of emotion regulation strategies described by Gross and colleagues (Gross & Thompson, 2007), including changing the evocative situation, re-directing attention, reframing thoughts about the situation, and direct physiological response modification through, for example, deep breathing. This subtest presents children vignettes, and they are asked to imagine they felt a certain undesirable emotion and what they could do to modify how they felt. A selection of strategies (e.g., "punch a pillow" or "take a few deep breaths") are presented for each scenario, some of which are more likely to be effective than others. We also included a brief self-report survey of self-control, which is presented after the direct assessment activities. Our goal in doing so was to supplement the emotional regulation knowledge task to enhance overall score reliability. We hypothesized that the emotion regulation knowledge (ERK) and self-reported self-control (SRSC) subtests would reflect a single underlying construct that we call "self-control."

The competencies SELweb LE is designed to assess are the targets of instruction in evidence-based SEL curricula (Weissberg et al., 2013). For example, most SEL curricula teach skills such as understanding other's thought and emotions, SPS, and self-management, all of which are competencies that SELweb LE is designed to assess. By using a predominantly performance-based direct assessment strategy, SELweb LE also measures competencies in ways that closely mirror specific skills that are targets of instruction. For example, many SEL curricula teach an approach to solving social problems that involves evaluating different possible solutions, picking the best one, and anticipating the consequences of that solution. Because SELweb LE presents children with hypothetical challenging situations, asks them to evaluate a range of responses, pick the best one, and predict what would happen next, children's SELweb LE responses on the SPS subtest closely mirror how they are learning in the context of SEL curricula.

1.3 | Assessment development process

To develop SELweb LE, we used an iterative process of item design, field testing, and item and subtest revision. During the first year of assessment development, a research team created items designed to reflect the competencies assessed by each subtest. Fifteen children completed the sample items while a member of the research team timed and observed the child. The researcher interviewed children about item format, clarity, and content in response to sample items for each subtest. Based on these early item tryouts, the team made minor modifications to directions

and item format. Based on time estimates, we then created enough items per subtest to produce an initial assessment that would take an estimated 40 min maximum. This included 12 items for a facial emotion recognition task, eight items each for CER, SPT, ERK, and six items for a SPS task, and 18 items for a delay discounting task. In addition, eight items were created to measure children's understanding of the relationship between thoughts, feelings, and behavior.

In an initial field test, we found that several subtests demonstrated low internal consistency reliability, including the thoughts, feelings, and behaviors task ($\alpha = .59$), ERK ($\alpha = .57$), and Perspective-Taking ($\alpha = .33$). Although the facial emotion recognition and delay discounting subtests yielded adequate internal consistency reliability, performance on those subtests were not associated with other variables. Based on these results, we dropped the facial emotion recognition, thoughts, feelings, and behaviors, and delay discounting tasks, and created new items for the SPT task. This paper presents data on the version of SELweb LE that resulted from that revision process.

1.4 | Study goals

This paper details the iterative process by which we evaluated SELweb LE (Study 1), replicated and expanded the field trial in a larger sample (Study 2), used the results of initial field trials to create additional items and alternate forms, and conducted a final field trial to evaluate the enhanced version's psychometric properties (Study 3). Study 1 evaluated SELweb LE's factor structure, score reliability, and evidence of convergent, discriminant, and criterion-related validity in a sample of 1,011 children. Study 2 replicated SELweb LE's factor structure and score reliability in a larger sample of 10,818 children and tested measurement equivalence across ethnicity, sex, and language. Study 3 evaluated the performance of an expanded item pool and revised number of items per subtest, again examining factor structure, score reliability, and criterion-related validity using additional validation measures in a sample of 3268 children. Together, these studies provide evidence of the reliability of SELweb LE scores, the validity of the assessment as reflected by factor structure and correlation with other measures, and measurement equivalence across demographic groups. We intended SELweb LE to be used by educators to characterize student competencies and measure student growth. Table 2 outlines standards of evidence for this intended use. Our goal in subjecting SELweb LE to empirical validation was to evaluate its performance in relation to these standards of evidence.

2 | STUDY 1

2.1 | Methods

2.1.1 | Sample and recruitment

Three school districts in the Chicago metropolitan area and one district in the Northeastern United States used SELweb LE to guide routine instructional practice. For this reason, the University's institutional review board (IRB) granted a waiver of informed consent to use de-identified SELweb LE data for research purposes. These districts self-administered SELweb LE to 1011 children in fourth through sixth grades. Almost all students (1007; 99.6%) had complete data. Because of the small amount of missing data, we omitted cases with any missing data. Sample characteristics are summarized in Table 3.

Parents of 384 children who completed SELweb LE consented to their children's participation in additional data collection. Of these 384 children, 199 were in third and fourth grade, which reflect the lower grade boundary for SELweb LE (fourth grade) and the upper grade boundary for one validation measure, SELweb EE (third grade). Reasoning that each assessment would maintain reasonable measurement properties in these boundary grades, we saw this as an opportunity to co-administer the two assessments to evaluate convergent and discriminant validity. Accordingly, those third and fourth graders whose parents consented to their participation completed SELweb EE, which served as

TABLE 2 Assessment goals and evidentiary criteria

Assessment goal	Evidentiary criteria	Criterion met?		
		Study 1	Study 2	Study 3
1. SELweb LE scores each reflect a specific competence that is distinct from other measured competencies.	CFA establishes hypothesized factor structure with CFI, IFI ≥ .95 and RMSEA ≤ .06.	Yes	Yes	Yes
	Each score is more associated with the score on an alternate measure of the same construct (convergent validity) than a measure of an adjacent construct (discriminant validity)	Partially (Convergent: 3 out of 3; Discriminant: 2 out of 3)	–	–
2. Total scores provide a consistent measure of competence.	Internal consistency ≥ .80, temporal stability ≥ .60	Partially (2 of 4 composite ICs = .78; all temporal stability > .60)	Partially (1 composite IC = .79; another = .75)	Partially (All IC ≥ .80; 3 out of 4 temporal stabilities > .60)
3. Scores can be interpreted as competencies that are meaningfully associated with youth behavior and achievement.	Significant associations between performance on SELweb EL and measures of student behavior, and achievement.	Yes	–	Yes
4. Total scores mean the same thing for boys and girls, for children from different ethnic groups, and for children who complete the assessment in English and in Spanish.	Multi-group CFA establishes configural, metric, and scalar invariance across sex, ethnicity, and language.	–	Yes	–

TABLE 3 Sample characteristics, english-speaking students

Characteristic	Study 1($n = 1011$)		Study 2($n = 10,818$)		Study 3($n = 3268$)	
	M	(SD)	M	(SD)	M	(SD)
Age	10.5	1.0	10.5	.8	10.3	.8
SSIS skills[a]	101.6	17.1	–	–	–	–
SSIS problem behavior[a]	97.6	15.0	–	–	–	–
SSIS academic	97.0	15.8	–	–	–	–
DESSA mini T	-	–	–	–	56.0	9.6
	n	%	n	%	n	%
Sex—Male	508	50.2	5,353	49.4	1,616	49.4
Ethnicity[b]						
White	241	23.8	1,917	17.7	1,519	46.5
Black	131	13.0	2,894	26.8	171	5.2
Hispanic	590	58.4	5,002	46.2	832	25.5
Asian	26	2.6	399	3.7	584	17.9
Other	23	2.3	459	4.2	162	5.0
Grade						
3	98	9.7	–	–	–	–
4	377	37.3	4,967	45.9	1,465	40.2
5	423	41.8	5,259	48.6	1,343	36.8
6	113	11.2	592	5.5	460	12.6
Setting						
Districts	4		12		9	
Schools	8		95		21	
Classrooms	68		491		153	

Note. This table shows demographic data on English-speaking students from all three studies.
[a]Estimated from the subset of students for whom data on student sex was available.
[b]147 students (1.4% of the sample) were missing ethnicity data.

our alternate measure of the same constructs. In addition, teachers completed behavior rating scales for 341 of the consented children in third through sixth grade. In terms of sample characteristics, there were no differences in the proportion of boys and girls who did and did not participate in the additional data collection. Compared to the larger sample from which they were recruited, a higher proportion of children in the add-on study were Black (22.6% vs. 9.1%), and a lower proportion were White (12.8% vs. 28.2%).

2.1.2 | Procedures

Educators administered SELweb LE to students in a group setting. Children logged in to SELweb LE and completed the assessment independently wearing headphones. All instructions were narrated within the program, and respondents had the option to have on-screen text read aloud by pressing a speaker icon. SELweb LE took approximately 35 min to complete. Additional data were collected from children participating in the add-on study through interviews at their school with trained members of the research team and teacher rating scales.

2.1.3 | Measures

SELweb LE. SELweb LE is a web-based assessment that includes five subtests made up of engaging interactive tasks designed to measure distinct components of SE competencies. Subtest descriptions and correspondence to CASEL competencies are listed below (see also Table 1).

For the CER subtest, children were presented with 12 illustrated and narrated stories about characters in situations that would invoke mixed or social emotions. Mixed emotion stories involve situations that plausibly cause two simultaneous emotions. Stories from the CER subtest involves ethical rule violations (guilt), publicly visible successes (pride), and publicly visible errors (embarrassment). Based on Pons et al. (2004), we developed scoring rules based on the content of the vignettes wherein each hypothesized correct selection was awarded one point. Because alternative responses are plausibly correct, we considered initial scoring rules as hypotheses subject to empirical tests as described below. Scoring rules were developed based on consensus among research team members about the emotions a situation would be likely to elicit.

The SPT subtest consists of three stories. Children are asked questions about the story characters' words, actions, or intentions. To answer questions correctly, children must make accurate inferences about story characters' mental state. Correct responses were assigned a score of one. Incorrect responses were assigned a score of zero.

For the SPS subtest, children are asked to imagine themselves in hypothetical challenging social situations involving, for example, peer entry, ambiguous provocation, and disagreements. Children are first asked about their goal preference ("What do you want to happen?"). Parallel to the construction of SELweb EE (McKown et al., 2016) and based on Crick and Dodge's (1996) findings about social goals, responses reflect the following categories of goals: relational (resolve the problem), avoidant (prevent escalation), or hostile (dominate or seek retribution). Next, children are asked for their solution preference ("What will you do?"). Response categories include positive-assertive, passive, or aggressive. Finally, children are asked to predict what would happen if they implemented their preferred solution ("If you [repeat chosen solution], what would happen?").

Scoring rules parallel those for SELweb EE (McKown et al., 2016). For goal preference, wanting to resolve the problem was awarded two points, wanting to avoid escalation was awarded one point, and wanting to dominate or get retribution was awarded zero points. For solution preference, positive-assertive responses were awarded two points, passive responses were awarded one point, and aggressive responses were awarded zero points. If a child selected a positive-assertive response, the predicted consequence was scored such that a positive resolution was assigned a score of one. Other responses received a score of zero.

For the ERK subtest, children are presented with 11 illustrated vignettes about people who feel disappointment, sadness, upset, or excessive excitement. They are asked what they could do to feel differently. Response options are based on the Gross and Thompson (2007) taxonomy of emotion regulation strategies and included strategies to change the situation provoking them, re-direct attention, reframe thoughts about the situation, and physiological response modification through, for example, deep breathing.

For each situation, children selected an emotion regulation strategy. Response options were constructed so that one of the responses was more likely to be effective than the others based on Gross and Thompson (2007). Distractors were designed to be plausible but ineffective—for example, punching a pillow to feel less upset about something. Correct responses were constructed such that they reflected an effective emotion regulation strategy as described by theory (Gross & Thompson, 2007). Competent emotion regulation strategies involve altering the situation (e.g., walking away from a situation), redirecting attention to (e.g., focusing on something that is not upsetting), reframing the situation (e.g., thinking about the situation in a way that is not upsetting), or direct physiological intervention (e.g., taking a few deep breaths). Those responses were awarded a score of one, and all other responses were scored zero. As was the case with CER, we considered these scoring rules to be hypotheses to be tested empirically as described in the Results and Discussion section.

The SRSC subtest is a seven-item self-report questionnaire in which children rate several statements about their self-control skills (e.g., "Even when someone is bothering me, I stay calm"). Respondents could have the items read aloud by clicking on them. Statements were rated on a five-point Likert scale (1 = Not at all true; 2 = A little true; 3 = Pretty true; 4 = Very true; 5 = Extremely true).

Alternate measure of SEL. SELweb EE was used as an alternate SE competence assessment to validate SELweb LE in grades 3 and 4. SELweb EE is a web-based direct assessment that measures facial emotion recognition, SPT, SPS, and self-control in grades K through 3(McKown et al., 2016). In pilot work, we found that SELweb EE yielded score reliabilities ≥.70 with fourth graders (1 year past the upper range of SELweb EE), and SELweb LE yielded score reliabilities ≥.70 with third graders (1 year below the lower range of SELweb LE). We therefore administered both measures only to children at these boundary grades. In this study, SELweb EE composite internal consistency reliabilities, calculated per procedures described by McKown et al. (2016), averaged .72.

Teacher reported social behavior. The teacher form of the Social Skills Improvement System (SSIS; Gresham & Elliott, 2008) was selected as a developmentally appropriate criterion measure establishing validity evidence by association with other variables, or criterion-related validity. The SSIS is a rating scale in which teachers rate the frequency of children's socially skilled and problem behavior. These behaviors are different from SE competencies measured by SELweb LE. However, if SELweb LE is indeed measuring competencies that shape social decisions, performance on SELweb LE should be associated with more frequent positive behavior and less frequent problem behavior. The SSIS was therefore administered via REDCap and used to analyze the relationship between SELweb LE performance and child behavior. For the SSIS, teachers indicated how frequently children demonstrated socially-skilled and problem behaviors and academic skills. Teachers were given a small monetary incentive for each SSIS form completed. Score reliabilities were as follows: Social Skills $\alpha = .98$, Problem Behaviors $\alpha = .96$, and Academic Competence $\alpha = .98$.

2.2 | Results and discussion

2.2.1 | Testing item scoring rules

Items on CER and ERK each had more than one plausible correct response. We therefore considered each item's initial scoring rule to be a scoring hypothesis. We used a variant of the "most popular option" method described by Guo et al. (2016), by which the most frequently selected option is scored as correct. Specifically, we assumed that options selected by respondents with the highest average overall score were correct. To test scoring hypotheses, we first computed a total score within each subtest by summing item scores using the initial scoring rule. For each item, we then computed the average overall subtest score of children who selected the hypothesized correct response and for children who selected hypothesized incorrect responses. Our scoring hypothesis was supported if children who selected the hypothesized correct response scored, on average, higher on the subtest overall than children who selected hypothesized incorrect responses. If children who selected another response also achieved higher average scores than children who selected alternatives, and that response reflected a plausible interpretation of the item content, we revised the scoring rule to give credit for that response.

For all CER items, children who selected the hypothesized correct response achieved higher average scores than children who selected other responses, supporting our scoring hypotheses. In addition, for three items, children who selected hypothesized incorrect responses that were nevertheless plausible achieved above-average overall scores. In those cases, scoring rules were revised so that those responses were awarded a score of one point.

For all ERK items, the average scores of children who selected hypothesized correct items were higher than the average scores of children who selected other responses, supporting our scoring hypotheses. As was the case for CER, for several ERK items, the average score of children who selected one of the distractor items was lower than the hypothesized correct response, but higher than the average scores of students who selected either of the remaining distractors. In these cases, when that response reflected an effective emotion regulation strategy, we awarded partial

credit. The final scoring system for ERK awarded two points for correct responses and one point for partially correct responses.

To evaluate the impact of revised scoring on the overall scores for CER and ERK, we computed total scores as the sum of item scores using both the original scoring system and the revised scoring system and computed the correlation between those scores. For CER, the correlation between total scores using the two scoring systems was $r = .95$. For ERK, the correlation between total scores using the two scoring systems was $r = .94$. These findings suggest that item scoring revisions did not have a substantial impact on overall score.

2.2.2 | Factor structure

For all three samples, to test multivariate normality, observed score skewness and kurtosis were computed and are summarized in Table 1_Supp. (see supporting material). For all scores, the absolute value of the skewness was less than 2. Kurtosis of 20 out of 21 variables had an absolute value less than 3. To evaluate the impact of these deviations from normality, Monte Carlo simulations with 200 bootstrap samples were run for each sample. For these analyses, Amos drew random, and therefore normally distributed, samples with the same means, variances, and covariances as the observed data. The distribution of parameter estimates from the simulated data were compared with parameter estimates from the observed data. There were no statistically significant differences between parameters estimated from bootstrap samples and those estimated from observed data. This suggests that deviations from normality did not have a meaningful impact on parameter estimates.

We used Amos (24.0.0, Arbuckle, 2008) to run confirmatory factor analyses with maximum likelihood estimation to evaluate the fit of SELweb LE scores to the hypothesized three-factor model. Consistent with recommended CFA reporting practices (Jackson et al., 2009), we included multiple fit indices in evaluating models, including the Incremental Fit Index (IFI), the Comparative Fit Index (CFI), and the Root Mean Squared Error of Approximation (RMSEA). These indicators are sensitive to model misspecification but not sample size. Consistent with the recommendations of Hu and Bentler (1999), we characterized a good fit of the model to the data as a CFI and IFI of at least .95, and a RMSEA less than .06.

SELweb LE scores fit the hypothesized three-factor model shown in Figure 1. One factor, which we labeled "understanding others," reflected the scores from CER and SPT. A second factor, which we labeled "social problem-solving," reflected scores for goal, solution, and consequences from the SPS subtest. A third factor, which we labeled "self-control," reflected scores from ERK and SRSC. We also tested a second-order model with an overall SE competence latent variable with three indicators: understanding others, SPS, and self-control. This model fit the data well (IFI = 1.00; CFI = 1.00; RMSEA = .032). This finding supports the use of an overall composite score reflecting the three SE competencies measured by SELweb LE. For each construct, to create a composite score, we standardized the observed scores that make up the composite, averaged those z-scores, and re-standardizing the resulting average. To create an overall score, we averaged the three composite scores and standardized the resulting average.

Recognizing the known limitations of modeling latent variables with two observed scores (Kline, 2015), we created item bins such that more than two observed scores loaded on each latent variable and re-ran analyses. Model fit and factor loadings were not meaningfully different from the simpler model described above, and so we maintained the simpler model.

2.2.3 | Score reliability

We computed the internal consistency (Cronbach's alpha) of each subtest (CER, SPT, SPS, ERK, and SRSC) using item scores within each subtest. To compute the reliability of composite scores described in the confirmatory factor models, we used procedures described by Nunnally and Bernstein (1994) which estimates composite reliability based on: (a)

Confirmatory Factor Model of SELweb LE
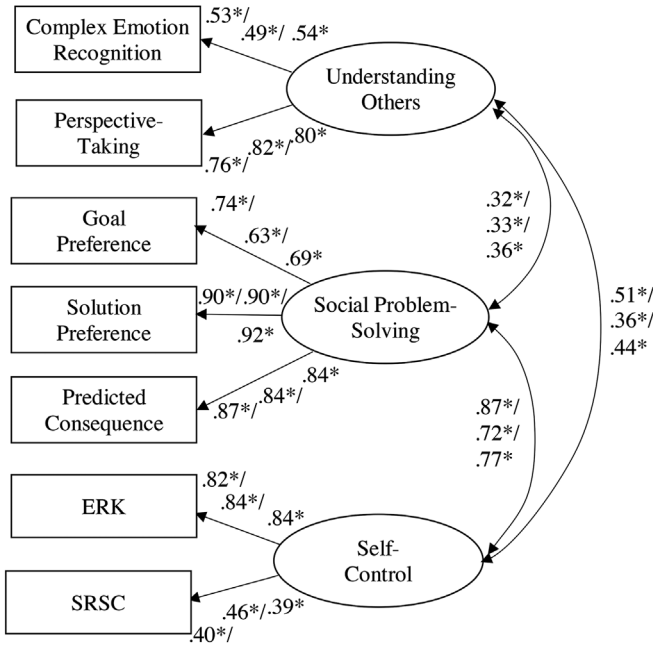
IFI = 1.00/.98/.98
CFI = 1.00/.98/.98
RMSEA = .025 (90% CI .000 - .046)/ .056 (90% CI .043 - .065)/ .059 (90% CI .054 - .063)

Note. All coefficients are standardized. Numbers before the "/" are from Study 1 data; numbers after the first "/" are from Study 2; numbers after the second "/" are from Study 3. For simplicity's sake, not all model parameters are shown in the figure.

FIGURE 1 Confirmatory Factor Model of SELweb LE. *Note.* All coefficients are standardized. Numbers before the "/" are from Study 1 data; numbers after the first "/" are from Study 2; numbers after the second "/" are from Study 3. For simplicity's sake, not all model parameters are shown in the figure

factor loadings of each observed score and its latent variable, (b) the reliability of each observed score, and (c) the covariance between observed scores, as follows:

$$r_{yy} = 1 - \frac{\Sigma b_i^2 \sigma_i^2 - \Sigma b_i^2 \sigma_i^2 r_{ii}}{\sigma_Y^2}$$

Where:

r_{yy} is the reliability of the latent variable

b_i is factor weights of each score i associated with the latent variable

r_{ii} is the reliability of each score i

σ_i^2 is the variance of each score i

σ_Y^2 is the summed variance of the obtained scores

Internal consistency and temporal reliabilities are summarized in Table 4 and show evidence of reliability in the scores of understanding others ($\alpha = .78$), SPS ($\alpha = .89$), self-control ($\alpha = .78$), and overall social emotional competence ($\alpha = .91$). As is always the case when indicators of a composite are correlated (Nunnally & Bernstein, 1994), the relia-

TABLE 4 SELweb LE internal consistency and temporal stability reliability, by factor and subtest

Factor and subtest	Study 1		Study 2		Study 3	
	α	r-r	α	r-r	α	r-r
SELweb LE—Factor scores						
Overall SEC	.91	.78	.90	–	.91	.71
Understanding others (UO)	.78	.74	.79	–	.80	.68
Social problem-solving (SPS)	.89	.66	.86	–	.82	.58
Self-control (SC)	.78	.65	.75	–	.80	.63
SELweb LE—Component scores						
UO–Complex emotion recognition (CER)	.64	.58	.66	–	.69	.51
UO–Social perspective-taking (SPT)	.73	.67	.72	–	.75	.67
SPS–Goal	.66	.53	.62	–	.55	.44
SPS–Solution	.74	.60	.68	–	.59	.52
SPS–Consequence	.75	.64	.70	–	.64	.54
SC–Emotion regulation knowledge (ERK)	.70	.64	.66	–	.71	.56
SC–Self-reported self-control (SRSC)	.80	.57	.81	–	.80	.62

Note. α = Cronbach's alpha; *r-r* = Six-month test-retest correlation.

bility of composites reflecting broad areas of SE competence were higher than the score reliabilities of the subscales from which they were derived, which ranged from $\alpha = .64$ to $\alpha = .80$. For applied interpretation, the composite scores are most appropriate to use, as scores for the broad composites have higher reliabilities than score for the indicators from which they are derived.

A subset of 563(57%) field trial participants completed SELweb LE a second time approximately 4 months after the first administration (Mean interval = 128 days, range = 103–138). To compute the temporal stability reliability of each subtest, we computed Pearson's correlations of scores from the first and second administrations. Temporal stability reliabilities of composite scores ranged from .65 to.78 and are summarized in Table 4.

These results suggest that SELweb LE yields reliable scores reflecting children's understanding of others' emotions and perspectives, SPS skill, and self-control. Reliability estimates met most of the evidentiary criteria outlined in Table 2. However, the understanding others internal consistency of .78 was .02 below the .80 standard.

2.2.4 | Associations with alternate measures of SE competence

Correlations between SELweb EE and SELweb LE scores are presented in Table 5. The pattern of findings supported the convergent and discriminant validity of the understanding others, SPS, and overall composite scores in grades three and four. Specifically, SELweb LE's understanding others score was most strongly correlated with SELweb EE's SPT score, which focuses on children's ability to infer others' perspectives. The understanding others score was not as strongly correlated with other SELweb EE scores, providing evidence of discriminant validity. The same pattern was present for SPS and the overall composite—each of these SELweb LE scores was most strongly associated with the corresponding SELweb EE score.

The SELweb LE self-control score was significantly associated with the SELweb EE self-control score. However, it was also associated with other SELweb EE scores, supporting its convergent but not discriminant validity. In particular, the magnitude of the association between the SELweb LE self-control score and the SELweb EE self-control score ($r = .25$) was less than the association between the SELweb LE self-control score and the SELweb EE SPS sore

TABLE 5 Correlations between SELweb LE scores and corresponding SELweb EE scores, study 1

	SELweb EE score				
SELweb LE composite score	SECoverall	Emotion recognition	Perspective-taking	Problem-solving	Self-control
SEC overall	**.54***	.15*	.41*	.41*	.35*
Understanding others	.32*	**.02***	**.41***	.06*	.32*
Problem-solving	.48*	.18*	.27*	**.51***	.26*
Self-control	.42*	.14*	.26*	.40*	**.25***

Note. Table shows Pearson's correlation coefficients. SEC = social emotional composite.
*p < .05.

TABLE 6 Association between SELweb LE performance and behavioral and academic outcomes

	SSIS[a]			DESSA[b]
Variable	Socialskill	Problem behavior	Academic competence	T score
Age	−.01	.02	.00	.18*
Sex	.31*	−.19*	.09	.44*
White	.32	−.14	.46*	.19
Black	.03	−.12	.31	−.43
Hispanic	.24	−.29	.19	.11
Asian	.96*	−.67*	.99*	.34*
SEC	.31*	−.32*	.40*	.24*

Note. Table shows standardized Pearson's correlation coefficients.
[a]Study 1.
[b]Study 2.
*p < .05.

(r = .40). This suggests that SELweb LE and the chosen criterion measure, SELweb EE, measure distinct dimensions of self-control. Specifically, SELweb LE emphasizes knowledge of emotion management. In contrast, SELweb EE uses game-like tasks to measure dimensions of inhibitory control including delay of gratification and frustration tolerance in response to real-time decision-making tasks. Furthermore, SELweb LE's vignette-based self-control subtest asks respondents to choose effective solutions to a specific class of problem (emotion dysregulation). This may be why the SELweb LE's self-control score is associated with SELweb EE's SPS score.

2.2.5 | Associations with teacher-reported social behavior

We analyze the association between the overall SELweb LE score and teacher report of social skills and problem behaviors on the SSIS, controlling for child age, sex, and ethnicity using hierarchical linear models (HLM) (HLM; Raudenbush & Bryk, 2002). As shown in Table 6, performance on SELweb LE was significantly associated in expected directions with teacher-reported social skills and problem behaviors. We also examined the association between each SELweb LE subscore and teacher report on the SSIS, controlling for child age, sex, and ethnicity. Each of the subscores was significantly associated with each SSIS score in the expected direction. In addition, we examined the association between the SELweb LE subscores entered simultaneously, and teacher report on the SSIS. In these analyses, understanding others

and self-control but not SPS were significantly and positively associated with social skills ($\beta = .13$ and $.21$, respectively) and academic competence ($\beta = .39$ and $.15$, respectively). Self-control but not understanding others or SPS was significantly and negatively associated with problem behaviors. These findings support the criterion-related validity of the overall SELweb LE score. It is noteworthy that overall performance on SELweb LE was most strongly associated with teacher-reported academic competence. Future research should examine whether this is because SELweb LE is measuring cognitive skills associated with academic performance, is measuring SE skills associated with academically engaged behaviors that teachers report on scales like the SSIS or originates from another source.

3 | STUDY 2

Study 2 replicated SELweb LE's factor structure and score reliability and tested measurement equivalence across ethnicity, sex, and language in a large independent sample.

3.1 | Method

3.1.1 | Recruitment and sample

The sample included children from six school districts in the Western, Midwestern, Southern, and Northeastern regions of the United States who completed SELweb LE during the 2018–2019 school year. Each measurement equivalence analysis included a different subset of children, depending on the availability of demographic information used to conduct the analyses. Ethnicity measurement equivalence analyses included all children who completed the English language version of SELweb LE. Sex measurement equivalence analyses focused on a subset of children whose data included information about child sex. Language measurement equivalence analyses included all children who completed SELweb LE in English or Spanish. Of the 10,818 students who began SELweb LE, 10,699(98.8%) completed all subtests. Because of the small amount of missing data, we omitted cases with any missing data. Sample characteristics are described in Table 3.

The ethnicity measurement equivalence sample analyses included 9781 children in fourth through sixth grades from 76 elementary schools in six school districts. Of those children, 9677(98.8) had complete SELweb LE data. The 119 cases (1.2%) with missing data were omitted from analyses. Only one of the six districts supplied information about child sex. Accordingly, the sex measurement equivalence sample included a subset of 1345 children in fourth through sixth grades from 12 elementary schools in one school district.

Compared to children who did not have information about sex, children who did have information about sex achieved statistically significant slightly higher average scores on complex emotion regulation recognition (14.0 vs. 13.8, $SD = 2.8$), SPT (.86 vs. .82, $SD = .14$), goal preference (1.38 vs. 1.34, $SD = .41$), and predicted consequence (.54 vs. .52, $SD = .27$). In addition, compared to children who did not have information about sex, children who had information about sex were more likely to be Asian (6.2% vs. 3.5%) and Black (29.0% vs. 27.3%). An additional 1367 fourth through sixth grade children from 69 elementary schools in six school districts completed the Spanish language version of SELweb LE. Of those children, 1342(98.0%) were missing no SELweb data. The 27 cases (2.0%) with missing data were omitted from analyses.

3.1.2 | Measures

Children in this sample completed the version of SELweb LE described in Study 1. At the time of administration, upon logging into SELweb LE, children selected whether they wished to complete the assessment in English or in Spanish.

3.2 | Results and discussion

3.2.1 | Confirmatory factor analysis

We replicated the CFA from Study 1. Fit statistics are presented with Figure 1.

3.2.2 | Score reliabilities

Score reliabilities were computed as described in Study 1 and are reported in Table 4.

3.2.3 | Measurement equivalence

Measurement equivalence was tested by comparing multi-group nested confirmatory factor analysis models with varying degrees of equality constraints (Dimitrov, 2010; Millsap, 2011). Grouping variables for measurement equivalence analyses included sex, ethnicity, and language. These analyses are based on the three-factor model described previously.

The most basic question about measurement equivalence is whether the factor structure is the same across groups (configural invariance). Configural invariance is tested by imposing no between-group constraints on model parameters and testing whether the factor structure fits the data equally well for all groups. Assuming configural invariance assumptions are met, a second important question is whether factor loadings are equivalent for different groups (metric invariance). Metric invariance means that a one unit change in the latent construct is reflected by the same change in the observed variables for all groups. Metric invariance is tested by imposing between-group equality constraints on factor loadings and again testing model fit. Metric invariance is demonstrated if, compared to the configural invariance model, the metric invariance model fit is not substantially reduced. Assuming metric invariance requirements are met, a third important question is whether latent intercepts are equivalent for different groups (scalar invariance). Scalar invariance means that at a given level of the latent variable, people from different groups achieve the same score on the observed variables. Scalar invariance is tested by imposing between-group equality constraints on latent intercepts (along with the constraints imposed in the metric invariance model). Scalar invariance is demonstrated if, compared to the metric invariance model, the scalar invariance model fit is not substantially reduced.

In all models tested below, the chi-square tests of overall model fit and differences between the fit of nested models were tested. The chi-square test of model fit is sensitive to sample size. Because of this, when samples are large, the chi-square test can indicate that the data do not fit the model even when the fit of the data to the model is excellent (Brannick, 1995; Ullman, 2006). Model fit was therefore also evaluated with CFI and RMSEA, both of which are less sensitive to sample size. Based on guidelines suggested by Dimitrov (2010), we interpreted the configural model as a good fit to the data when CFI ≥ .95 and RMSEA ≤ .06. Metric invariance models for each grouping were compared to the configural model, and scalar invariance models were compared to their respective metric invariance model. Using guideline suggested by Chen (2007), metric and scalar invariance were supported if the change in model fit from the less restrictive model to the more restrictive model was CFI < .01 and RMSEA < .015.

In this sample, there were small but statistically significant age differences between children from different ethnic groups and between children who completed the assessment in English and Spanish. To account for these differences, for the ethnicity and language measurement equivalence analyses, age was entered as a predictor of all observed scores. Measurement equivalence analyses are summarized in Table 7. Based on the criteria described above, ethnicity, sex, and language fit configural, metric, and scalar invariance models.

TABLE 7 Measurement invariance fit statistics for the three-factor model, study 2 data

Model	df	Δdf	χ^2	$\Delta\chi^2$	RMSEA	ΔRMSEA	CFI	ΔCFI
Ethnicity invariance								
Configural	55	–	492.8*	–	.029	–	.980	–
Metric	71	16	542.7*	49.9*	.026	−.003	.979	.001
Scalar	99	28	582.1*	49.4*	.023	−.003	.978	.001
Language invariance								
Configural	22	–	451.6*	–	.042	–	.983	–
Metric	26	4	455.4	3.8	.039	−.003	.983	.000
Scalar	33	7	469.1	13.7	.035	−.004	.983	.000
Sex invariance								
Configural	22	–	461.2*	-	.044	–	.982	–
Metric	26	4	464.5	3.3	.040	−.004	.982	.000
Scalar	33	7	572.7*	108.2*	.040	.000	.978	.004

*$p < .05$.

3.3 | Revision plan

In Study 1 and Study 2, at the level of composite scores, score reliabilities were variable, with internal consistency reliabilities under .80 for understanding others and self-control, and in the high .80's for SPS. Our goal was to achieve internal consistency reliabilities of .80 or above for the three main composites and .90 or above for the overall composite. That would require an improvement in score reliability for understanding others and emotion management. To improve these scores' reliabilities, we used two strategies—first, for ERK and Complex Emotion Recognition, we revised or replaced a small number of items that demonstrated the lowest item-total correlations. Second, for ERK, we added one item to the scale.

For SPS, the composite reliability of .89 was substantially greater than our target score reliability of .80. We viewed having more items than necessary as a potential hindrance to usability. Reducing the number of vignettes presented in this subtest would reduce the overall length of the assessment while yielding an acceptable score reliability. Accordingly, we used the Spearman-Brown prophecy formula (Spearman, 1910) to estimate the number of vignettes on this subtest that would yield a score reliability of .80. We estimated that a reduction from eight to six vignettes would accomplish this aim and proceeded accordingly.

In addition, we wished to create alternate forms of SELweb LE. To that end, we generated additional items for all subtests except for SRSC. We designed each new item to parallel an existing item, with modified content. For example, in vignettes, the characters, settings, or situations might be modified. We assigned items to four forms such that each item was present in at least two forms.

4 | STUDY 3

The purpose of Study 3 was to replicate analyses focused on factor structure, score reliability, and criterion-related validity in a version of SELweb LE with revised items, as described in the Study 2 Revision Plan.

4.1 | Methods

4.1.1 | Recruitment and sample

A total of 3268 fourth to sixth grade children from 21 schools in nine school districts from the Midwest, West, and Northeast participated in the study and used SELweb LE as part of routine instructional practice. We also collected de-identified criterion measure data already collected in one district to assess criterion-related validity. Demographic characteristics of the whole sample and the subsamples with criterion measures are described in Table 3. Of the 3268 children who began SELweb LE, 3237(99.1%) had complete data. Because of the small amount of missing data, we omitted cases with any missing data.

4.1.2 | Measures

Children in this sample completed the revised version of SELweb LE.

One district also administered the Devereux Student Strengths Assessment, mini version (DESSA-mini; LeBuffe et al., 2018), an eight-item teacher rating scale designed to measure teacher perceptions of child SE competence. Like our use of the teacher report on the SSIS, the DESSA-mini provides a measure of social behavior by which to evaluate the criterion-related validity of SELweb LE. Internal consistency of the DESSA-mini was $\alpha = .93$. The district provided T scores based on national norms.

4.2 | Results and discussion

4.2.1 | Scoring

To score each subtest, we totaled individual item scores within each scale. To place scores for each subtest on a common scale across forms, we standardized the scores within form.

4.2.2 | Factor structure

We replicated the CFA from Studies 1 and 2. Fit statistics are presented with Figure 1.

4.2.3 | Score reliability

We computed internal consistency reliability separately for each of the four versions of SELweb LE and averaged those values, which are presented in Table 4. Composite score reliabilities and temporal stability reliabilities were computed using procedures described in Study 1. Consistent with our goals, revisions of the assessment resulted in composite internal consistencies equal to or greater than .80.

A subset of 681(21%) participants completed SELweb LE a second time approximately 4 months after the first administration (Mean interval = 136 days, range = 111–177). To compute the temporal stability reliability of each subtest, we computed Pearson's correlations of scores from the first and second administrations. Temporal stability reliabilities of composites ranged from .58 to.71 and are summarized in Table 4.

4.2.4 | Criterion-related validity

To assess criterion-related validity, we used HLM to assess the relationship between SELweb LE performance and teacher-reported behavior from the DESSA-mini, controlling for child age, sex, and ethnicity. As shown in Table 6, we found a significant association between SELweb LE performance and the DESSA-mini T score.

4.2.5 | Form equivalence

Finally, we evaluated the extent to which the four different forms of SELweb LE, each with different item sets, functioned the same. To that end, we conducted measurement equivalence analyses, as described in Study 2, with "test form" as the grouping variable. The models fit the assumptions of configural, metric, and scalar invariance.

The results of Study 3 suggest that SELweb LE yields reliable scores reflecting children's understanding of others' emotions and perspectives, SPS skill, and self-control. Furthermore, Study 3 CFA results support structural validity. Finally, criterion-related validity of SELweb LE using a criterion measure distinct from those used in Study 1.

5 | GENERAL DISCUSSION

5.1 | Summary of key findings

The purpose of SELweb LE is to characterize SE competence levels in children across fourth through sixth grades. Denham (2018) proposed that assessment developers attend to guiding theoretical frameworks, attend to developmental trajectories in both item design and context, align assessment content with SEL standards, and develop tools with the stakeholder and user experience in mind. Furthermore, a 2019 Hanover Research report acknowledges that performance tasks, or direct assessments completed by children are an emerging area of assessment for which more work is needed. SELweb LE was designed with these considerations in mind. Through iterative design, field-testing, and revision, we developed a developmentally-appropriate, theoretically-informed measure of SE competencies reflected in state standards and evidence-based SEL curricula. Furthermore, the three studies described here provide evidence that SELweb LE meets evidentiary criteria for its intended uses.

The first criterion for SELweb LE's intended use is to yield scores that reflect competencies that are distinct from one another. Confirmatory factor analyses yielded a strong fit to a three-factor structure that included three partially independent scores, one reflecting children's understanding of others' thoughts and feelings, another reflecting SPS skill, and a third reflecting self-control. The understanding others score was modestly correlated with SPS and self-control scores, which was consistent with the conclusion that these are partially independent competencies. That SPS and self-control scores were more highly correlated is consistent with research reporting associations between executive functions, self-regulation, and social functioning in clinical populations (e.g., van Nieuwenhuijzen & Vriens, 2012; van Nieuwenhuijzen et al., 2017; van Rest et al., 2019; Wolfe et al., 2015). In addition, the SPS items and the emotional regulation knowledge items resembled one another in an important way: they both ask what can be done in in response to a challenging situation. The strong association between scores on these subtests may reflect common method variance. Together these data suggest that the ability to understand others' thoughts and feelings, as measured by SELweb LE, is distinct from problem-solving and self-control, while problem-solving and self-control are mores strongly related.

Furthermore, understanding others and SPS scores demonstrated evidence of both convergent and discriminant validity. The SELweb LE self-control score, while associated with the alternate measure of self-control and therefore showing evidence of convergent validity, was also correlated with other SE competencies measured by the alternate

measure, suggesting that it may be capturing other competencies apart from self-control. It will be important to continue to probe the empirical proximity of SE competencies to one another.

The second criterion for SELweb's intended use is that scores provide consistent measures of competencies. In studies 1 and 2, two of four internal consistency reliability estimates were slightly below the .80 criterion set for the intended use argument. Revisions to item content and number, reflected in Study 3, resulted in all scores meeting conditions of satisfaction for all internal consistency scores and three of four temporal stability reliability scores. It is important to note that score reliabilities—both internal consistency and temporal stability—met the evidentiary criteria for our intended use for composite scores, but not the indicator scores that contributed to those composites.

The third criterion for SELweb LE's intended use was that scores reflect competencies that are meaningful, as judged by an association with indicators of behavior. Data from Study 1 and Study 3 demonstrated significant associations in the expected direction between performance on SELweb LE and teacher-reported outcomes, including socially skilled behavior, problem behaviors, and academic functioning. These findings suggest that performance on SELweb LE reflects competencies that are consequential as reflected by the association between SELweb LE performance and indicators of important areas of functioning.

The final criterion for SELweb LE's intended use was that scores mean the same thing for children from different demographic groups. Multi-group CFA tests of measurement equivalence supported configural, metric, and scalar invariance across sex, ethnicity, and language form. These results suggest that scores derived from SELweb LE have comparable meaning for children from diverse backgrounds.

5.2 | Contributions to theory and practice

These studies have implications for theory. Many models of SE competence have been proposed (Berg et al, 2017; Blyth et al., 2018). Studying and clarifying the nature of any psychological phenomenon requires strong measurement tools. SELweb LE reflects a theoretically-derived and developmentally-appropriate tool for studying the nature and development of SE competencies in middle childhood. It is particularly well-suited to advancing knowledge about the CASEL model. Factor analytic findings across all three studies are consistent with and partially support the broad CASEL model of SEL. Specifically, the understanding others score reflects CASEL's "social awareness," the self-control score reflects CASEL's "self-management," and the SPS score reflects CASEL's "relationship skills" and "responsible decision-making." While it is not feasible to assess all dimensions of SE competence, SELweb LE measures important dimensions of the competencies described by CASEL.

These studies also have implications for practice. Many educators are implementing SEL practices that are based on the CASEL model. Furthermore, educators have expressed a desire to use assessment to understand and address student needs. To do so requires assessments that characterize student level of SE competence in ways that correspond to the competencies educators intend to nurture in their students. Because SELweb LE uses a predominantly direct assessment approach to measure CASEL competencies, it is well-suited to providing educators with information they may use to decide what to teach to whom. Future research should specifically examine how educators can best use SE competence assessment data to make instructional decisions that have a positive impact on student outcomes.

5.3 | Limitations and future directions

In developing SELweb LE, we recognized that it is not possible to feasibly measure every SE competence in one assessment and instead used the literature to guide our choice of competencies to measure. SELweb LE touches on four of the five CASEL competencies. No subtests measure "self-awareness," and the subtests measure some but not all the components of the remaining four CASEL competencies. Future work should focus on developing measurement strategies to assess SE competencies not currently measured by SELweb LE—either in brief assessment

formats or another multidimensional assessment that would complement the existing suite. One of the inherent tensions in so doing will be balancing comprehensive construct coverage with feasible assessment length. It may be most constructive to identify components of CASEL's five competence areas that are most consequential and focus measurement development on those "high-leverage" competencies. Surely other combinations are possible and future work should examine these possibilities. Future assessments, for example, should include indicators of responsible decision-making along with a measure of self-awareness to provide a more complete test of the CASEL model.

Performance on SELweb LE and teacher ratings of child behavior were consistently associated, but the magnitude of the association was modest. If both forms of assessment are designed to measure SE competencies, why is the magnitude of the association not greater? Direct assessment largely measures mental processes that children apply in social contexts. Teacher ratings measure observable behavior. On its face, these are two different dimensions of children's social repertoire. Direct assessments and rating scales are not redundant. Rather, they are complementary and together may provide a fuller picture of student SE competence than either form of assessment used alone.

Although SELweb LE demonstrated evidence of measurement equivalence across sex, ethnicity, and language form, we note two areas for future research. First, despite evidence of measurement equivalence, possible cultural differences in interpretation and response to items merit future study (Hecht & Shin, 2015). For example, children from different cultures may interpret complex mixed and social emotions and the situations that cause them differently; in addition, in some cultures, it may be more appropriate to enlist an adult to help than to assert oneself (Ellis et al., 2017; Günsoy et al., 2015). Second, it is important to note that the competencies SELweb LE does measure may be more socially consequential for children from some demographic groups than others and other SE competencies that SELweb LE is not designed to measure, such as "code switching," or the ability to fluidly alter responses to situations based on changes in the cultural context, may be highly consequential. It will be important to continue to clarify the range of SE competencies relevant to diverse learners and to devise fair assessments of those competencies.

We have argued that direct assessment provides insights into children's competence that are otherwise difficult to ascertain. However, it is important to note that one of SELweb LE's subtests is a self-report survey. In addition, children's responses to hypothetical vignettes are not the same thing as their responses to real-life situations. Indeed, the association between performance on SELweb LE and teacher rating scales suggests that while there is a moderate association between student performance on SELweb LE and behavior, what SELweb LE is measuring is different from observable behavior. It may be useful to consider SELweb LE as a measure of what children know, while behavior rating scales measure what children do.

Scoring rules for some subtests were developed based on theory and confirmed empirically. In a small number of cases, scoring rules were modified based on data. This method acknowledges some uncertainty around the values associated with some responses. Two findings lend some confidence that these scoring rules are durable across samples and contexts. First, correlations between scores derived from the hypothesized scoring system and scores derived from the revised scoring system were well above $r = .90$. Second, final scoring rules from our initial studies were applied in studies two and three and score reliabilities and factor structure remained consistent. Future research examining item characteristics of subtests with these scoring rules will help confirm the equivalence of scoring rules across samples and contexts.

Assessment can and should play a key role in supporting high-quality practice and its improvement as programs scale (Fagan, et al., 2019; McKown et al., 2021). If practitioners can characterize student SE competencies to guide instruction and monitor the quality with which they were implementing SEL programs, data on student competencies and implementation quality can be used to support high-quality instruction that targets student needs. Currently, with some exceptions, SE assessment and curricula operate in parallel universes. Future research should examine how SE competence assessment, alongside measures of practice, can support a growing field to maintain quality at scale.

6 | CONCLUSION

SEL policies, programs, and practices are proliferating rapidly. In this context, the field needs high-quality, scalable, feasible SE competence assessments to guide theory and practice. Building on prior work, SELweb LE is designed to fulfill a need for a technically sound, usable, feasible, informative, assessment that measures SE competencies supported by theories and relevant to SEL instructional programs and state standards in the late elementary grades.

DATA AVAILABILITY STATEMENT

No data are available.

REFERENCES

1. Arbuckle, J. L. (2008). *Amos (version 24.0.0)*. SPSS, Inc.
2. Assessment Work Group (2019). *Student social and emotional competence assessment: The current state of the field and a vision for its future.* Collaborative for Academic Social and Emotional Learning.
3. Atwell, M. N., & Bridgeland, J. M. (2019). *Ready to lead: A 2019 update of principals' perspectives on how social and emotional learning can prepare children and transform schools.* Civic. https://files.eric.ed.gov/fulltext/ED602977.pdf
4. Berg, J., Osher, D., Same, M. R., Nolan, E., Benson, D., & Jacobs, N. (2017). *Identifying, defining, and measuring social and emotional competencies: Final report.* American Institutes for Research. https://www.air.org/sites/default/files/downloads/report/Identifying-Defining-and-Measuring-Social-and-Emotional-Competencies-December-2017-rev.pdf
5. Bernard, H. R., Killworth, P., Kronenfeld, D., & Sailer, L. (1984). The problem of informant accuracy: The validity of retrospective data. *Annual Review of Anthropology, 36,* 162–180.
6. Bitsakou, P., Antrop, I., Wiersema, J. R., & Sonuga-Barke, E. J. S. (2006). Probing the limits of delay intolerance: Preliminary young adult data from the Delay Frustration Task (DeFT). *Journal of Neuroscience Methods, 151*(1), 38–44. https://doi.org/10.1016/j.jneumeth.2005.06.031
7. Blyth, D. A., Jones, S., & Borowski, T. (2018). *SEL frameworks: What are they and why are they important?* Establishing Practical Social-Emotional Competence Assessment Work Group. https://measuringsel.casel.org/wp-content/uploads/2018/09/Frameworks-A.1.pdf
8. Brannick, M. T. (1995). Critical comments on applying covariance structure modeling. *Journal of Organizational Behavior, 16*(3), 201–213. https://doi.org/10.1002/job.4030160303
9. Chen, F. F. (2007). Sensitivity of goodness of fit indexes to lack of measurement invariance. *Structural Equation Modeling: A Multidisciplinary Journal, 14*(3), 464–504. https://doi.org/10.1080/10705510701301834
10. Crick, N. R., & Dodge, K. A. (1994). A review and reformulation of social information-processing mechanisms in children's social adjustment. *Psychological Bulletin, 115,* 74–101. https://doi.org/10.1037/0033-2909.115.1.74
11. Crick, N. R., & Dodge, K. A. (1996). Social information-processing mechanisms in reactive and proactive aggression. *Child Development, 67*(3), 993–1002. https://doi.org/10.2307/1131875
12. Crowne, D. P., & Marlowe, D. (1960). A new scale of social desirability independent of psychopathology. *Journal of Consulting Psychology, 24*(4), 349–354. https://doi.org/10.1037/h0047358
13. Denham, S. A. (1986). Social cognition, prosocial behavior, and emotion in preschoolers: Contextual validation. *Child Development, 57*(1), 194–201. https://doi.org/10.2307/1130651
14. Denham, S. A. (2018). *Keeping SEL developmental: The importance of a developmental lens for fostering and assessing SEL competencies.* Collaborative for Academic, Social, and Emotional Learning. https://measuringsel.casel.org/wp-content/uploads/2018/11/Frameworks-DevSEL.pdf
15. Denham, S. A., Bassett, H. H., Mincic, M., Kalb, S., Way, E., Wyatt, T., & Segal, Y. (2012). Social-emotional learning profiles of preschoolers' early school success: A person-centered approach. *Learning and Individual Differences, 22*(2), 178–189. https://doi.org/10.1016/j.lindif.2011.05.001
16. Devine, R. T., & Hughes, C. (2013). Silent films and strange stories: Theory of mind, gender, and social experiences in middle childhood. *Child Development, 84,* 989–1003. https://doi.org/10.1111/cdev.12017
17. Diazgranados, S., Selman, R. L., & Dionne, M. (2016). Acts of social perspective taking: A functional construct and the validation of a performance measure for early adolescents. *Social Development, 25*(3), 572–601. https://doi.org/10.1111/sode.12157
18. Dimitrov, D. M. (2010). Testing for Factorial Invariance in the Context of Construct Validation. *Measurement and Evaluation in Counseling and Development, 43*(2), 121–149. https://doi.org/10.1177/0748175610373459
19. DiPerna, J. C., & Elliott, S. N. (2002). Promoting academic enablers to improve student achievement: An introduction to the mini-series. *School Psychology Review, 31*(3), 293–297. https://doi.org/10.1080/02796015.2002.12086156

20. Dubow, E. F., Tisak, J., Causey, D., Hryshko, A., & Reid, G. (1991). A two-year longitudinal study of stressful life events, social support, and social problem-solving skills: Contributions to children's behavioral and academic adjustment. *Child Development*, *62*(3), 583–599. https://doi.org/10.1111/j.1467-8624.1991.tb01554.x

21. Duckworth, A. L., & Kern, M. L. (2011). A meta-analysis of the convergent validity of self-control measures. *Journal of Research in Personality*, *45*(3), 259–268. https://doi.org/10.1016/j.jrp.2011.02.004

22. Duckworth, A. L., & Yeager, D. S. (2015). Measurement matters: Assessing personal qualities other than cognitive ability for educational purposes. *Educational Researcher*, *44*, 237–251. https://doi.org/10.3102/0013189X15584327

23. Durlak, J. A., Weissberg, R. P., Dymnicki, A. B., Taylor, R. D., & Schellinger, K. B. (2011). The impact of enhancing students' social and emotional learning: A meta-analysis of school-based universal interventions. *Child Development*, *82*(1), 405–432. https://doi.org/10.1111/j.1467-8624.2010.01564.x

24. Dusenbury, L., Dermody, C., & Weissberg, R. P. (2018). *State scorecard scan*. Collaborative for Academic Social and Emotional Learning. https://casel.org/wp-content/uploads/2018/09/csi-scorecard-sept2018.pdf

25. Dweck, C. S., Chiu, C. Y., & Hong, Y. Y. (1995). Implicit theories and their role in judgments and reactions: A word from two perspectives. *Psychological Inquiry*, *6*(4), 267–285. https://doi.org/10.1207/s15327965pli0604_1

26. Ellis, B. J., Bianchi, J., Griskevicius, V., & Frankenhuis, W. E. (2017). Beyond risk and protective factors: An adaptation-based approach to resilience. *Perspectives on Psychological Science*, *12*(4), 561–587. https://doi.org/10.1177/1745691617693054. Epub 2017 Jul 6. PMID: 28679332.

27. Fagan, A. A., Bumbarger, B. K., Barth, R. P., Bradshaw, C. P., Cooper, B. R., Supplee, L. H., & Walker, D. K. (2019). Scaling up evidence-based interventions in US public systems to prevent behavioral health problems: Challenges and opportunities. *Prevention Science*, *20*(8), 1147–1168. https://doi.org/10.1007/s11121-019-01048-8

28. Fan, X., Miller, B. C., Park, K. E., Winward, B. W., Christensen, M., Grotevant, H. D., & Tai, R. H. (2006). An exploratory study about inaccuracy and invalidity in adolescent self-reports. *Field Methods*, *18*, 223–244. https://doi.org/10.1177/152822X06289161

29. Gresham, F. M., & Elliott, S. N. (2008). *Social skills improvement system: Rating scales manual*. NCS Pearson.

30. Gross, J. J., & Thompson, R. A. (2007). Emotion regulation: Conceptual foundations. In (J. J. Gross Ed.), *Handbook of emotion regulation* (pp. 3–24). Guilford Press.

31. Günsoy, C., Cross, S. E., Uskul, A. K., Adams, G., & Gercek-Swing, B. (2015). Avoid or fight back? Cultural differences in responses to conflict and the role of collectivism, honor, and enemy perception. *Journal of Cross-Cultural Psychology*, *46*(8), 1081–1102. https://doi.org/10.1177/0022022115594252

32. Guo, H., Zu, J., Kyllonen, P., & Schmitt, N. (2016). Evaluation of different scoring rules for a noncognitive test in development. *ETS Research Report Series*, *2016*(1), 1–13. https://doi.org/10.1002/ets2.12089

33. Hamilton, L. S., Doss, C. J., & Steiner, E. D. (2019). *Teacher and principal perspectives on social and emotional learning in America's schools: Findings from the American Educator Panels*. RAND Corporation. https://www.rand.org/pubs/research_reports/RR2991.html

34. Happé, F. (1994). An advanced test of theory of mind: Understanding of story characters thoughts and feelings by able autistic, mentally handicapped, and normal children and adults. *Journal of Autism and Developmental Disabilities*, *24*, 129–154. https://doi.org/10.1007/BF02172093

35. Hecht, M. L., & Shin, Y. (2015). Culture and social and emotional competencies. In (J. A. Durlak, C. E. Domitrovich, R. P. Weissberg, & T. P. Gullotta Eds.), *Handbook of social and emotional learning: Research and practice* (pp. 50–64). The Guilford Press.

36. Hu, L. T., & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling*, *6*(1), 1–55. https://doi.org/10.1080/10705519909540118

37. Innovation Research and Training, Inc. (2008). *The Social Information Processing Application (SIP-AP) [computer software]*. Innovation Research and Training, Inc.

38. Jackson, D. L., Gillaspy, J. A., & Purc-Stephenson, R. (2009). Reporting practices in confirmatory factor analysis: An overview and some recommendations. *Psychological Methods*, *14*(1), 6–23. https://doi.org/10.1037/a0014694

39. Kline, R. P. (2015). *Principles and practice of structural equation modeling, fourth edition*. The Guilford Press.

40. Korkman, M., Kirk, U., & Kemp, S. (2007). *NEPSY II: Clinical and interpretive manual*. Harcourt Assessment, PsychCorp.

41. Kuntsi, J., Stevenson, J., Oosterlaan, J., & Sonuga-Barke, E. (2001). Test–retest reliability of a new delay aversion task and executive function measures. *British Journal of Developmental Psychology*, *19*(3), 339–348. https://doi.org/10.1348/026151001166137

42. Kupersmidt, J. B., Stelter, R., & Dodge, K. A. (2011). Development and validation of the social information processing application: A web-based measure of social information processing patterns in elementary school-age boys. *Psychological Assessment*, *23*(4), 834–847. https://doi.org/10.1037/a0023621

43. LeBuffe, P. A., Shapiro, V. B., & Robitaille, J. L. (2018). The Devereux Student Strengths Assessment (DESSA) comprehensive system: Screening, assessing, planning, and monitoring. *Journal of Applied Developmental Psychology*, *55*, 62–70. https://doi.org/10.1016/j.appdev.2017.05.002

44. Mayer, J. D., Caruso, D. R., & Salovey, P. (2005). *The Mayer-Salovey-Caruso emotional intelligence test-Youth version (MSCEIT-YV), research version*. Multi Health Systems.

45. McKown, C., Kharitonova, M., & Russo-Ponsaran, N. M. (2021). Development and validation of a shortened form of SEL-web EE, a web-based assessment of children's social-emotional comprehension. Assessment, advance online publication. https://journals.sagepub.com/doi/abs/10.1177/10731911211046044

46. McKown, C., Russo-Ponsaran, N. M., Allen, A. A., Johnson, J., & Russo, J. (2016). Web-based assessment of children's social-emotional comprehension. *Journal of Psychoeducational Assessment*, *34*, 322–338. https://doi.org/10.1177/0734282915604564

47. Millsap, R. E. (2011). *Statistical approaches to measurement invariance*. Routledge.

48. Nelson, D. A., & Crick, N. R. (1999). Rose-colored glasses: Examining the social information-processing of prosocial young adolescents. *The Journal of Early Adolescence*, *19*(1), 17–38. https://doi.org/10.1177/0272431699019001002

49. Nowicki, S., & Duke, M. P. (1994). Individual differences in the nonverbal communication of affect: The diagnostic analysis of nonverbal accuracy scale. *Journal of Nonverbal Behavior*, *18*(1), 9–35. https://doi.org/10.1007/BF02169077

50. Nunnally, J. C., & Bernstein, I. (1994). *Psychometric theory*. McGraw-Hill.

51. Oberle, E. (2018). Early adolescents' emotional well-being in the classroom: The role of personal and contextual assets. *Journal of School Health*, *88*(2), 101–111. https://doi.org/10.1111/josh.12585

52. Payton, J., Weissberg, R. P., Durlak, J. A., Dymnicki, A. B., Taylor, R. D., Schellinger, K. B., & Pachan, M. (2008). The positive impact of social and emotional learning for kindergarten to eighth-grade students: Findings from three scientific reviews. *Collaborative for Academic, Social, and Emotional Learning*, https://files.eric.ed.gov/fulltext/ED505370.pdf

53. Peters, C., Kranzler, J. H., & Rossen, E. (2009). Validity of the Mayer-Salovey-Caruso emotional intelligence test: Youth version—research edition. *Canadian Journal of School Psychology*, *24*(1), 76–81. https://doi.org/10.1177/0829573508329822

54. Pons, F., Harris, P. L., & de Rosnay, M. (2004). Emotion comprehension between 3 and 11 years: Developmental periods and hierarchical organization. *European Journal of Developmental Psychology*, *1*(2), 127–152. https://doi.org/10.1080/17405620344000022

55. Raudenbush, S. W., & Bryk, A. S. (2002). *Hierarchical linear models: Applications and data analysis methods (2nd ed.)*. Sage Publications, Inc.

56. Rivers, S. E., Brackett, M. A., Reyes, M. R., Mayer, J. D., Caruso, D. R., & Salovey, P. (2012). Measuring emotional intelligence in early adolescence with the MSCEIT-YV: Psychometric properties and relationship with academic performance and psychosocial functioning. *Journal of Psychoeducational Assessment*, *30*(4), 344–366. https://doi.org/10.1177/0734282912449443

57. Rotheram, M. J. (1987). Children's social and academic competence. *The Journal of Educational Research*, *80*(4), 206–211. https://doi.org/10.1080/00220671.1987.10885753

58. Russo-Ponsaran, N. M., McKown, C., Karls, A., & Wu, I. Y. H. (2021). Psychometric properties of virtual environment for social information processing, a social information processing simulation assessment for children. *Social Development*, *30*(3), 615–640. https://doi.org/10.1111/sode.12512

59. Salovey, P., & Mayer, J. D. (1990). Emotional intelligence. *Imagination. Cognition, and Personality*, *9*(3), 185–211. https://doi.org/10.2190/DUGG-P24E-52WK-6CDG

60. Selman, R. L. (1980). *The growth of interpersonal understanding: Developmental and clinical analyses*. Academic Press.

61. Shrauger, J. S., & Osberg, T. M. (1981). The relative accuracy of self-predictions and judgments by others in psychological assessment. *Psychological Bulletin*, *90*(2), 322–351. https://doi.org/10.1037/0033-2909.90.2.322

62. Smith, R. L., & Rose, A. J. (2011). The "cost of caring" in youths' friendships: Considering associations among social perspective taking, co-rumination, and empathetic distress. *Developmental Psychology*, *47*(6), 1792–1803. https://doi.org/10.1037/a0025309

63. Spearman, C. (1910). Correlation calculated from faulty data. *British Journal of Psychology, 1904-1920*, *3*(3), 271–295. https://doi.org/10.1111/j.2044-8295.1910.tb00206.x

64. Ullman, J. B. (2006). Structural equation modeling: Reviewing the basics and moving forward. *Journal of Personality Assessment*, *87*(1), 35–50. https://doi.org/10.1207/s15327752jpa8701_03

65. van Nieuwenhuijzen, M., Van Rest, M. M., Embregts, P. J. C. M., Vriens, A., Oostermeijer, S., Van Bokhoven, I., & Matthys, W. (2017). Executive functions and social information processing in adolescents with severe behavior problems. *Child Neuropsychology*, *23*(2), 228–241. https://doi.org/10.1080/09297049.2015.1108396

66. van Nieuwenhuijzen, M., & Vriens, A. (2012). (Social) Cognitive skills and social information processing in children with mild to borderline intellectual disabilities. *Research in Developmental Disabilities*, *33*(2), 426–434. https://doi.org/10.1016/j.ridd.2011.09.025

67. van Rest, M. M., Matthys, W., Van Nieuwenhuijzen, M., De Moor, M. H. M., Vriens, A., & Schuengel, C. (2019). Social information processing skills link executive functions to aggression in adolescents with mild to borderline intellectual disability. *Child Neuropsychology*, *25*(5), 573–598. https://doi.org/10.1080/09297049.2018.1495186

68. Weissberg, R. P., Goren, P., Domitrovich, C., & Dusenbury, L. (2013). *2013 CASEL guide: Effective social and emotional learning programs–Preschool and elementary school edition.* Collaborative for Academic, Social, and Emotional Learning. https://files.eric.ed.gov/fulltext/ED581699.pdf

69. Willoughby, M., Kupersmidt, J., Voegler-Lee, M., & Bryant, D. (2011). Contributions of hot and cool self-regulation to preschool disruptive behavior and academic achievement. *Developmental Neuropsychology*, *36*(2), 162–180. https://doi.org/10.1080/87565641.2010.549980

70. Wolfe, K. R., Vannatta, K., Nelin, M. A., & Yeates, K. O. (2015). Executive functions, social information processing, and social adjustment in young children born with very low birth weight. *Child Neuropsychology*, *21*(1), 41–54. https://doi.org/10.1080/09297049.2013.866217

SUPPORTING INFORMATION

Additional supporting information can be found online in the Supporting Information section at the end of this article.