JOURNAL ON EDUCATION IN EMERGENCIES

Teachers' Observations of Learners' Social and Emotional Learning: Psychometric Evidence for Program Evaluation in Education in Emergencies

Author(s): Ha Yeon Kim, Kalina Gjicali, Zezhen Wu, and Carly Tubbs Dolan

Source: *Journal on Education in Emergencies*, Vol. 7, No. 2 (December 2021), pp. 57-103

Published by: Inter-agency Network for Education in Emergencies

Stable URL: http://hdl.handle.net/2451/63538

DOI: https://doi.org/10.33682/3nr1-3ksq

REFERENCES:

TEACHERS' OBSERVATIONS OF LEARNERS' SOCIAL AND EMOTIONAL LEARNING: PSYCHOMETRIC EVIDENCE FOR PROGRAM EVALUATION IN EDUCATION IN EMERGENCIES

Ha Yeon Kim, Kalina Gjicali, Zezhen Wu, and Carly Tubbs Dolan

ABSTRACT

Rigorous evaluation of social and emotional learning programs requires the use of measures that provide reliable and valid information on the meaningful differences in children's social emotional skills across treatment and control groups, as well as changes over time. In contexts affected by conflict and crisis, few measures can provide the evidence required to support their use in program evaluations, which limits stakeholders' ability to determine whether a program is working, how well it is working, and for whom. The Teachers' Observation of Learners' Social Emotional Learning, known as the TOOLSEL, holds promise for addressing this gap. The TOOLSEL is a teacher-report questionnaire about children's behavior as observed in natural classroom settings. It is used to assess a set of social, emotional, behavioral, and cognitive competencies among primary school-age children in fragile, conflict-affected settings. In this article, using the data from a sample of 3,661 Syrian refugee children who were enrolled in formal Lebanese public schools and had access to a nonformal remedial support program, we report evidence on the psychometric soundness of the TOOLSEL. We provide empirical evidence of the TOOLSEL's reliability and validity, and that the TOOLSEL captured these Syrian refugee children's social and emotional learning skills in ways that were unbiased and comparable across treatment groups, gender, age, and time. We also provide recommendations for using the TOOLSEL, including ways to improve its feasibility, reliability, and validity.

Received March 14, 2020; revised November 30, 2020, and April 21, 2021; accepted August 2, 2021; electronically published December 2021.

Journal on Education in Emergencies, Vol. 7, No. 2

ISSN 2518-6833

INTRODUCTION

Diverse stakeholders are increasingly investing in the implementation of social and emotional learning (SEL) programs in humanitarian contexts (UNESCO 2018). SEL programs provide safe, predictable learning environments for conflict-affected children that can promote the social and emotional skills that are critical in bolstering their resilience, addressing risks proactively, and building competencies at scale (Betancourt et al. 2013; Burde et al. 2017; Jordans, Pigott, and Tol 2016). These skills are important developmental indices, and they promote better academic outcomes (Durlak et al. 2011), as well as labor market attainment and wellbeing over the longer term (Heckman, Stixrud, and Urzua 2006; Jones, Greenberg, and Crowley 2015). However, little rigorous research has been conducted on the impact SEL programming has on refugee children living in humanitarian contexts, which leaves a critical knowledge gap when making programmatic decisions about how to support conflict-affected children most effectively (UNESCO 2018; Bakrania et al. 2021).

Building the evidence base on SEL in humanitarian contexts requires having field-feasible measures of children's social and emotional skills that are psychometrically sound, fit for program-evaluation purposes, and appropriate for the context in which they are being implemented (Tubbs Dolan and Caires 2020). Historically, many measures of social and emotional skills have been adopted from existing tools and used "off-the-shelf" in crisis contexts, with little consideration of their intended purpose (e.g., screening test, formative assessment, program evaluation) or whether they can provide reliable, valid information about the target population and context (Tubbs Dolan 2017). However, merely translating a tool designed for a different culture and context into a new language does not guarantee that it will provide a valid measurement of SEL in a new context. At a minimum, stakeholders must assess the psychometric properties of existing measures when they are used in a new context or with a new population (AERA, APA, and NCME 2014).

This study is one attempt to generate evidence on the reliability and validity of a measure assembled from existing measures used in humanitarian contexts. The Teachers' Observation of Learners' Social and Emotional Learning, known as TOOLSEL, is designed to capture teachers' perceptions of primary school-age children's social, emotional, behavioral, and cognitive skills. It was specifically developed to evaluate an SEL program that targets these skills in nonformal education settings for Syrian refugee children living in Lebanon. In this article, we present the data we used from a large randomized controlled trial to provide psychometric evidence of TOOLSEL's effectiveness with these children.

BUILDING SOCIAL AND EMOTIONAL COMPETENCIES IN EDUCATION IN EMERGENCIES

Education programming in emergency contexts can provide children with a safe space and a structured routine that creates a sense of normalcy, as well as opportunities to develop supportive relationships and attain meaningful learning outcomes (UNESCO 2018; Davies and Talbot 2008). However, children in education in emergency (EiE) settings may enter their classrooms with psychosocial challenges stemming from their experiences of violence, forced migration, and exploitation, as well as myriad daily stressors (Betancourt et al. 2013; Burde et al. 2015), all of which can interfere with their ability to learn and to connect with their teachers and classmates (Burde et al. 2017; Kim et al. 2020). Given research suggesting that children in crisis settings may be at particular risk for difficulties with social and self-regulation skills, practitioners working in emergency contexts have targeted these skills as key components of SEL programs, such as the Better Learning Program (Shah 2017) and Five-Component SEL (Kim et al. 2021).

TOOLSEL was designed to address the need for measures that can be used in EiE classrooms to assess the status and improvement of such SEL skills reliably and validly. It captures a range of cognitive, emotional, and behavioral competencies that are hypothesized to be important for children's successful social and academic adaptation in classrooms in EiE settings, which teachers can observe through daily classroom interactions. TOOLSEL focuses specifically on capturing several important social competencies and challenges that children display in classroom environments, as well as the self-regulatory functions necessary for learning, such as executive function, and emotional and behavioral regulation. We briefly discuss these competencies below.

Classrooms—both physical classrooms in formal schools and other nonformal peer-group learning spaces—are a primary setting where many school-age children who have access to education are able to develop and maintain relationships. Research in non-EiE contexts has found that successful social adjustment—as indicated by positive social interactions such as prosocial behavior and peer acceptance—is related to concurrent and future academic outcomes (Furrer and Skinner 2003), and to social competence, emotional health, and positive school behaviors (Hartup 1996). On the other hand, social difficulties indicated by aggression, peer rejection, and victimization put children at increased risk of maladaptive social-emotional functioning, both in the present and over time (Gest, Welsh, and Domitrovich 2005; LaFontana and Cillessen 2002).

Self-regulation is another of the SEL skills relevant to and observable in classroom settings. Self-regulation involves a complex system of cognitive, emotional, and behavioral processes that inhibit or modulate children's predominant responses to stimuli, and that enable them to display more adaptive emotions and behaviors (Eisenberg, Smith, and Spinrad 2011; Rothbart and Rueda 2005). Indeed, US studies suggest that self-regulation is critical for children's ability to develop successful social relationships (Kochanska, Murray, and Harlan 2000) and academic competence (Raver et al. 2011). A recent study conducted with Syrian refugee children living in Lebanon (Kim et al. 2020), which used measures that were tested for reliability and validity with the sample, also confirmed that children's cognitive and behavioral regulation skills are predictive of their academic performance.

The cognitive aspects of self-regulation skills are often represented as executive-function skills, which refers to a broad set of cognitive capacities, including working memory (i.e., the ability to keep in mind goal-relevant information) and inhibitory control (i.e., the ability to stop oneself from performing a prepotent response; Blair and Razza 2007). Extensive research suggests that executive function is a key mechanism for children's self-regulation in school, which is foundational to their learning and school success (Hughes and Ensor 2011; Jacob and Parkinson 2015). Regulation of emotions is another aspect of self-regulation that represents the capacity to regulate one's emotions and behavior in order to produce adaptive responses to the demands of a situation (Rothbart and Rueda 2005). Evidence from non-EiE contexts suggests that regulation of emotions is related to children's academic success (Boekaerts and Pekrun 2015), and to their social competence and peer acceptance (Valiente et al. 2011). Lastly, behavioral regulation—that is, the capacity to modulate behavior to achieve a specific goal—is a third foundational skill that enables children to adjust and learn successfully in classroom settings (Duncan, McClelland, and Acock 2017).

MEASURING THE IMPACT OF SEL PROGRAMMING ON SOCIAL AND EMOTIONAL SKILLS IN EIE SETTINGS

Evaluating the impact of SEL programs on children's social and emotional skills in EiE settings requires measures that are field feasible and have strong evidence of psychometric soundness.

Field Feasibility

Using teacher rating measures, such as TOOLSEL, in an EiE context has several advantages in terms of feasibility, including that teachers' reports (1) are based on

accumulated knowledge of a particular child in various social and academic settings over a period of time, as compared to observation-based assessments that rely on a small set of short observation sessions; (2) are less likely to be subject to social-desirability bias or be dependent on children's self-awareness skills, as compared to self-report measures that require children to reflect and respond objectively about their own thoughts, feelings, and behaviors (Van de Mortel 2008); and (3) are low cost and easy to incorporate into the platforms commonly used for monitoring and evaluation, unlike interview protocols and performance-based measures that are expensive to develop and adapt, and that require lengthy data collection on individual children. While performance- or observation-based measures hold promise for measuring task- and context-specific skills and performance (Taylor et al. 2018), the cost to develop measures and collect data that are appropriate to a particular context and population may be prohibitive.

Psychometric Criteria

For a measure to be suitable for evaluation purposes, it must meet several psychometric criteria (Tubbs Dolan and Caires 2020). First, measures used for program-evaluation purposes must have strong evidence of coherence by consistently providing information on the unique and meaningful constructs the measures are intended to capture. Second, data from program-evaluation measures must be highly reliable, as an error in the data can attenuate the ability to determine the impact of a program (Raudenbush and Sadoff 2008). Third, data from program-evaluation measures should provide evidence that the measures function and that they capture the same SEL skills of children from different subgroups (e.g., of different gender and age groups) and over time, in order to assess differences by group and changes in the same set of skills. This criterion is known as measurement invariance. Fourth, measures developed to evaluate impact should be sensitive to program-induced change that may occur during the program. Lastly, the measure should capture the key behaviors of social, emotional, and cognitive skills by providing evidence of expected relations in terms of direction—that is, whether they are positively or negatively related—and of magnitude, relative to other theoretically related variables.

Potential Correlates of Teacher-Rated SEL Skills

A variety of factors beyond the skills themselves are likely to be related to teachers' ratings of children's SEL skills. These include characteristics such as age and gender, similar or related social and emotional skills, and experiences reported by other sources.

First, as children mature, they build the capacity to regulate their emotions and behavior (Cole, Michel, and Teti 1994), become aware of others' perspectives in a social situation and display more prosocial behaviors (Fabes and Eisenberg 1998), and become able to sustain their attention for longer periods of time (Lumley et al. 2002). Research suggests that children become better with age at planning their actions and controlling their impulses (Zelazo, Carlson, and Kesek 2008). As they develop (Zimmermann and Iwanski 2014), children also gradually develop adaptive emotional and behavioral regulation strategies.

Second, gender differences in social-emotional skills and behaviors are prominent across domains. A meta-analysis of gender differences in children's prosocial behavior confirms that girls generally exhibit more prosocial behavior than boys (Fabes and Eisenberg 1998). Evidence from studies with war-affected children is consistent with findings from those in non-EiE contexts, with teachers rating girls lower than boys in aggression and higher in prosocial behaviors (Elzein and Ammar 2010; Keresteš 2006). Research has found that boys tend to exhibit more problems paying attention and more disruptive behavior disorders than girls (Lumley et al. 2002). However, such differences could be blurred in cultural contexts where culture-specific beliefs, values, and gender stereotypes appear to be different (Brody 2000) and different measurement methods are considered (McRae et al. 2008).

Lastly, teachers' rating of students' SEL in classrooms is likely to be modestly correlated with similar concepts where different measures were used by different reporters. For example, social competence and prosocial behavior are expected to be negatively related to self-reports of bullying and victimization experienced in school, whereas social problems are likely to be positively correlated with victimization (Ellis et al. 2016). In addition, executive function measured using performance-based assessments would likely be related to teachers reports of children's working memory and classroom behaviors related to inhibitory control. Observer reports of behavioral regulation are also likely to be related to teachers' ratings of behavioral regulation.

While typically not highly correlated with performance- and observation-based or child self-report measures (Buckley and Krachman 2016), teachers' reports provide meaningful information, as their perception and interpretation of children's behavior can affect their interaction with the children and the children's outcomes (McKown and Weinstein 2008). Ultimately, examining the divergence and convergence of different measurement methods provides multifaceted information that is valuable in understanding children's social and emotional development in emergency contexts (De Los Reyes et al. 2015).

CURRENT STUDY

This study utilizes data collected from Syrian refugee children in nonformal education classrooms in Lebanon—a typical education setting in EiE contexts—and examines the psychometric properties of TOOLSEL, a teacher-report measure of children's SEL, in order to provide evidence of the tool's validity and reliability. We first provide evidence of the measure's internal coherence by identifying unique SEL constructs captured through the nonformal education teachers' perspectives on the TOOLSEL and report the internal consistency of the items for each construct. Then we test whether these SEL constructs are consistently measured across treatment groups, different age groups and genders, and across time (fall to spring). We next examine whether the SEL constructs differ by changes occurring during the programming period, by age, and by gender. Finally, we test the hypothesized association between the SEL constructs captured with TOOLSEL and the children's experience of victimization at school, behavioral regulation, and executive function, which are measured using different tools.

METHODS

Participants

We utilize data from a sample of Syrian refugee children living in Lebanon who were enrolled in nonformal remedial support programs; the data were collected as a part of a large, randomized controlled trial. During the 2017-2018 school year, the International Rescue Committee delivered nonformal remedial tutoring programming that was infused with SEL principles to Syrian refugee children in Lebanon's Bekaa and Akkar regions. The program was offered in community sites located close to the area where a large number of the Syrian refugees reside, either in spaces rented in buildings in urban/residential areas or in tent schools and classrooms built for the program in the informal settlement communities located in more rural areas. The parents or guardians of all participants provided written consent for their children to participate in the research. The participants included 3,661 students ages 5 to 16 (M=9.38, SD=2.27; 50% female) who were enrolled in grades 1 to 7 in Lebanese public schools; they came from 169 classrooms in the 57 community sites. At the time of the study, the children had been living in Lebanon an average of four years (M=4.13, SD=1.50), and the majority of them (86%) had not reported any interruption in their schooling. Students in 29 sites were

randomly assigned to a treatment condition, where an additional SEL intervention was implemented as a part of the tutoring programming. All programming was offered in Arabic. Data were collected in the fall at the beginning of the program (November: n=3,254) and at the end in the spring (May: n=2,952).

Measures

All items of each measure used in this study were translated from English into Arabic. They were adapted through rounds of iterative feedback from the International Rescue Committee's local practitioners, who were working closely with teachers and students in Lebanon to ensure an adequate linguistic, cultural, and contextual fit.

TOOLSEL

Given the scarcity of SEL measures developed locally with the Syrian refugee population, TOOLSEL is assembled from various teacher-report surveys of children's classroom behaviors that were developed and tested in the US The TOOLSEL items are drawn from three measures: the Teacher Observation of Child Adaptation-Checklist (TOCA-C; Koth, Bradshaw, and Leaf 2009); the Social Competence Scale (SCS; Conduct Problems Prevention Research Group 1990); and the Classroom Executive Function Survey (CEFS; Jones, Bailey, and Barnes 2015). Each item was rated on a five-point Likert scale ranging from 1="Never" to 5="Almost always." See Table A1 in the Appendix for the full list of items.

TOCA-C (Koth et al. 2009) is a teacher-report checklist, originally developed in the US to assess the social adaptive classroom behaviors of first-grade students as viewed and defined by their teachers. Selected items from the Prosocial Behavior, Concentration Problems, and Disruptive Behavior subscale were included in TOOLSEL. In studies in the US (Koth et al. 2009) and Greece (Kourkounasiou and Skordilis 2014), internal consistency was high for each of the subscales, with Cronbach's alpha ranging from 0.87 to 0.97.

TOOLSEL also includes items from the Emotion Regulation subscale of the SCS, which was originally created for the Fast Track Project (Conduct Problems Prevention Research Group 1990). Lastly, eight items from the CEFS (Jones et al. 2015) were included to capture teachers' perceptions of students' executive function skills. CEFS was specifically designed to measure children's demonstrated working memory, inhibitory control, and attention skills; it has been used previously in the EiE context, including in Lebanon.

Victimization experience in public schools

The school victimization experience was captured via a six-item questionnaire that asked children to reflect on their experience in public schools in the previous two weeks. The questions included the four items of the Victim subscale in the Illinois Bully Scale (Espelage and Holt 2001; e.g., "Other students pick on me." "I got hit and pushed by other students."), and an additional two items to reflect receiving harsh treatment from adults in school; this was common among the Syrian refugee children attending the public schools, according to anecdotal reports from the partner organization field practitioners ("Teachers, school directors, or other adults in public school pinched, pulled hair, or pulled ears." "Teachers, school directors, or other adults in public school hit me with an object such as a ruler, stick, or *tuyau* [PVC pipe]." Responses were measured on a scale of 0="Not at all." to 4="Absolutely yes." Internal consistency reliability was α=0.75 in the fall.

Preschool self-regulation assessment: Assessor report

Children's behavior regulation was rated by assessors using a 13-item version of the Preschool Self-Regulation Assessment: Assessor Report (PSRA-AR; Smith-Donald et al. 2007) adapted for a study in Zambia (McCoy et al. 2017). The PSRA-AR was originally designed to include assessors' ratings of each child's behavior as displayed during the performance-based PSRA assessment (e.g., "Pays attention to instructions and demonstration." "Remains in seat appropriately during test."). Each item was scored on a four-point Likert scale, with higher scores indicating better behavioral regulation.

Rapid assessment of cognitive and emotional regulation

The Rapid Assessment of Cognitive and Emotional Regulation (RACER; Ford et al. 2019) was used to assess two aspects of executive function, working memory and inhibitory control, on a random half of the current sample. RACER demonstrated good accuracy and reliability in testing in Peru (Hamoudi and Sheridan 2015), Lebanon, and Niger (Ford et al. 2018), and also was used in Ghana, Bangladesh, and Ethiopia. Working memory was measured using a Spatial Delayed Match to Sample Task (Goldman-Rakic 1996). Inhibitory control was measured using a Simon Task (Simon and Rudell 1967).

Analytic Approach

When using a measure in a new context and with a new population, conducting an empirical assessment of the psychometric properties is a necessary first step toward developing a locally developed and/or contextualized measure (AERA, APA, and NCME 2014). To do this, we conducted the analyses described below.

All descriptive analyses for this study were conducted using Stata SE15.1, and all factor analyses were conducted using Mplus 8.3 (Muthén and Muthén 2014).[1] First, to identify the unique SEL constructs underlying the TOOLSEL items, we identified and confirmed the TOOLSEL factor structure by conducting exploratory factor analyses (EFA) and confirmatory factor analyses (CFAs) at each time point (fall and spring). All items in the measurement models were estimated using weighted least squares mean and a variance adjusted estimator with a probit-link function (Lei 2009). The following criteria were used to assess the models' goodness of fit (Hu and Bentler 1999): RMSEA<0.08; CFI/ TLI>0.9; and SRMR<0.08.

Second, to assess internal consistency, we report Cronbach's α and McDonald's ω (Hayes and Coutts 2020; McDonald 1999) of each latent factor; ω does not assume equal factor loadings (i.e., all items contribute equally to measuring the construct of interest) and therefore is a better estimate of internal consistency than the conventional α (Revelle and Zinbarg 2009). While there are no definitive and universal guidelines, α>0.70 and ω>0.80 are generally considered acceptable/ highly reliable (Catalán 2019).

Third, we tested measurement invariance across the treatment and control groups, different age groups, and gender groups for each time point, and longitudinal invariance across time. Measurement invariance refers to the extent to which a set of items measures an underlying construct of interest in the same way across groups or time (Reise, Widaman, and Pugh 1993). This is done by testing the equivalence of (a) the factor structure in treatment, gender, and age groups, and across timepoints (configural invariance) to evaluate whether and to what extent the same latent constructs could be identified by the same manifest observations across groups and time points; (b) the factor loadings of the items across groups/ timepoints (metric invariance) in order to test whether the psychological meanings

1 To account for nested data structure where teachers reported on all individual children's SEL, all analyses were conducted using robust standard errors, adjusted for clustering at the classroom level. In all factor-analysis models, missing data at the item level were pairwise deleted (i.e., all available information was used from all cases) to preserve the full sample (Asparouhov and Muthén 2010).

of the measured latent constructs are equivalent across groups and time points; and (c) the item intercepts or thresholds across groups/timepoints (scalar invariance) to evaluate whether the means of different groups or observations at different time points can be compared on the same scale (Vandenberg and Lance 2000).[2]

Fourth, we tested hypothesized differences of the TOOLSEL constructs across treatment groups, age groups, gender groups, and assessment time (fall to spring) by comparing the intercept of the latent factors in the measurement invariance models. For example, to compare male and female students, we report intercepts of the latent factors for females in the scalar invariance model of the gender invariance analysis, where male students' mean is fixed at zero. And, lastly, we examined the extent of the measurement validity of TOOLSEL by investigating (a) the bivariate association of the TOOLSEL constructs across time; (b) the bivariate associations with other related constructs; and (c) partial correlations controlling for child demographic characteristics (age, grade, gender) using the ordinary least squares regression approach.

RESULTS

Identifying TOOLSEL Constructs

Exploratory Factor Analysis

Given the poor model fit of the five-factor confirmatory factor analysis models that reflect the original subscales the items came from (Table B1 in the Appendix), a series of exploratory factor analyses was used to conduct an empirically based exploration of the factor structure. All 28 items were included in the initial EFA models (see Table A1 for a full list of items and the items that were removed; see descriptive statistics of all items in Table A2). A four-factor solution consisting of 23 items was chosen due to the acceptable model fit and consistent patterns in the factor structure across the fall and the spring (Table B2). A list of items for the four subscales identified from the EFA are presented in Table 1.

2 The relative fit of each of these models was assessed against the configural model using criteria suggested by Chen (2007); metric invariance: ΔCFI<0.01, ΔRMSEA<0.015, ΔSRMR<0.030; scalar invariance: ΔCFI<0.01, ΔRMSEA<0.015, ΔSRMR<0.010.

Table 1: TOOLSEL Items by Subscales

Number	Construct	Item Code and Description
1	Prosocial Behavior and Academic Engagement	TOC1: In the last two weeks [your child]: Concentrates
2		TOC2: In the last two weeks [your child]: Is friendly
3		TOC3: In the last two weeks [your child]: Pays attention
4		TOC7: In the last two weeks [your child]: Works hard
5		TOC5: In the last two weeks [your child]: Is liked by classmates
6		TOC9: In the last two weeks [your child]: Shows empathy & compassion for other's feelings
7	Social Problems	TOC10: In the last two weeks [your child]: Gets angry when provoked by other children
8		TOC15: In the last two weeks [your child]: Fights
9		TOC12: In the last two weeks [your child]: Yells at others
10		TOC14: In the last two weeks [your child]: Is rejected by classmates
11		TOC20: In the last two weeks [your child]: Teases classmates
12	Working Memory Functioning	TOC21: In the last two weeks [your child]: Learns up to ability
13		CEFS1: In the last two weeks [your child]: Remembers lists or items in the correct order
14		CEFS2: In the last two weeks [your child]: Follows multiple-step instructions
15		CEFS3: In the last two weeks [your child]: Uses multiple rules to complete a task

Number	Construct	Item Code and Description
16	Emotional and Behavioral Regulation	CEFS4: In the last two weeks [your child]: Waits to be called on before responding
17		SCS11: In the last two weeks [your child]: Can calm down when excited or all wound up
18		CEFS6: In the last two weeks [your child]: Transitions easily to new activities, tasks, or major parts of the day (e.g., from recess)
19		CEFS8: In the last two weeks [your child]: Uses self-control techniques
20		SCS12: In the last two weeks [your child]: Can wait in line patiently when necessary
21		CEFS9: In the last two weeks [your child]: Waits patiently for her/his turn
22		CEFS10: In the last two weeks [your child]: Uses listening skills
23		SCS18: In the last two weeks [your child]: Controls temper when there is a disagreement

Note: Full set of items included in the initial analysis is available in Appendix A. Items labeled starting with TOC are taken from TOCA-C, with original item numbers used in TOCA-C. Similarly, items labeled starting with SCS were taken from SCS Emotional Regulation Scale, and items labeled starting with CEFS were taken from CEFS.

Confirmatory Factor Analysis

CFA with the four factors extracted from the EFA was run with the fall data and then modified to include two additional residual covariances (Table 2, Figure 1). This same final model obtained from the fall was tested with the endline (spring) data and yielded a result with an acceptable model fit (Table B3). All items loaded onto their respective factors with high factor loadings at $\lambda>0.50$. The final factor structure revealed that the TOOLSEL constructs represented a considerable departure from the original subscales. Specifically, Factor 1: Prosocial Behavior and Academic Engagement, was a combination of the positively worded items from the Prosocial Behavior and Concentration Problems subscales of TOCA-C. Factor 2: Social Problems consisted of items from the Disruptive Behavior and negatively worded items from the Prosocial Behavior subscales of the TOCA-C. Factor 3: Working Memory Functioning was composed of one item from the Concentration Problem subscale from the TOCA-C, "Learn up to ability," and three items from the CEFS that described the children's working memory capacity. Lastly, Factor 4: Emotional and Behavioral Regulation, consisted of three items from the SCS Emotion Regulation subscale and five items from CEFS that describe children's ability to inhibit impulsive behaviors and to participate in classroom activities. The final model allowed two sets of item covariance for Factor 4 for a better model fit, based on conceptual similarity: (a) items CEFS4, "Waits to be called on," and SCS11, "Can calm down when excited," and (b) items SCS12, "Can wait in line patiently," and CEFS9, "Waits patiently for turn." See Table 3 for the factor loadings of each item in both the fall and the spring. These four latent factors of teacher-reported SEL skills were highly correlated to each other, ranging from r=-0.453 to 0.877 in the fall, and from r=-0.351 to 0.889 in the spring (Figure 1).

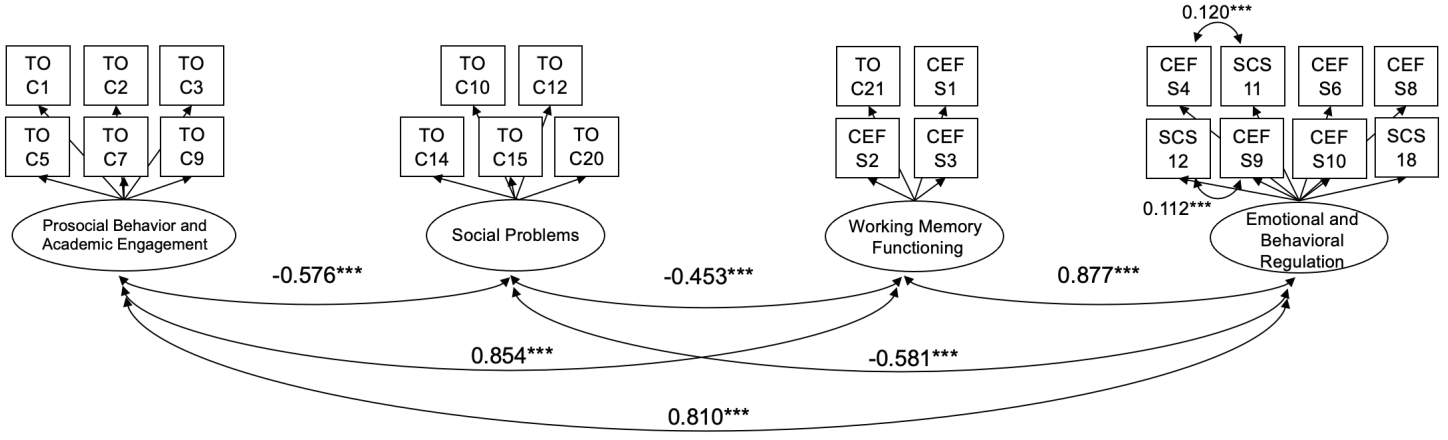
Table 2: Factor Loadings of the TOOLSEL at Fall and Spring from the Confirmatory Factor Analysis Final Model

		Fall			Spring		
		b	*SE*	*p*	*b*	*SE*	*p*
Prosocial Behavior and Academic Engagement (Fall α=0.921, =0.945; Spring α=0.932, =0.950)							
1	TOC1: Concentrates	0.913	0.007	0	0.932	0.006	0
2	TOC2: Is friendly	0.883	0.009	0	0.905	0.008	0
3	TOC3: Pays attention	0.903	0.008	0	0.905	0.008	0
4	TOC7: Works hard	0.778	0.014	0	0.805	0.013	0
5	TOC5: Is liked by classmates	0.882	0.009	0	0.896	0.008	0
6	TOC9: Shows empathy & compassion	0.781	0.015	0	0.805	0.014	0
Social Problems (Fall α=0.847, =0.900; Spring α=0.847, =0.886)							
1	TOC10: Gets angry when provoked	0.647	0.024	0	0.560	0.032	0
2	TOC15: Fights	0.875	0.014	0	0.879	0.014	0
3	TOC12: Yells at others	0.847	0.021	0	0.864	0.021	0
4	TOC14: Is rejected by classmates	0.892	0.014	0	0.892	0.014	0
5	TOC20: Teases classmates	0.714	0.022	0	0.737	0.020	0
Working Memory Functioning (Fall α=0.877, =0.909; Spring α=0.910, =0.928)							
1	TOC21: Learns up to ability	0.709	0.018	0	0.804	0.016	0
2	CEFS1: Remembers lists or items	0.851	0.009	0	0.899	0.009	0
3	CEFS2: Follows multistep instructions	0.901	0.009	0	0.927	0.007	0
4	CEFS3: Uses multiple rules	0.883	0.009	0	0.905	0.009	0

		Fall			Spring		
		b	*SE*	*p*	*b*	*SE*	*p*
Emotional and Behavioral Regulation (Fall α=0.960, =0.972; Spring α=0.964,; =0.973)							
1	CEFS4: Waits to be called on	0.881	0.009	0	0.911	0.007	0
2	SCS11: Can calm down when excited	0.872	0.009	0	0.897	0.009	0
3	CEFS6: Transitions easily to new activities	0.901	0.007	0	0.925	0.006	0
4	CEFS8: Uses self-control techniques	0.915	0.006	0	0.927	0.007	0
5	SCS12: Can wait in line patiently	0.909	0.007	0	0.924	0.007	0
6	CEFS9: Waits patiently for turn	0.905	0.008	0	0.919	0.006	0
7	CEFS10: Uses listening skills	0.919	0.007	0	0.928	0.007	0
8	SCS18: Controls temper	0.864	0.01	0	0.840	0.013	0

Note: Items labeled starting with TOC are taken from TOCA-C, with original item numbers used in TOCA-C. Similarly, items labeled starting with SCS is taken from SCS Emotional Regulation Scale, and items labeled starting with CEFS were taken from CEFS.

Figure 1: Factor-Structure Diagrams Displaying Model Parameters at Fall (top) and Spring (bottom)



Internal Consistency of Subscales

Table 3 also presents Cronbach's alpha estimates for scores from the empirically derived TOOLSEL subscales. All subscales have high internal reliability, ranging from α=*0*.85-0.96 to ω=*0*.87-0.97.

Measurement Invariance

Using the final, empirically derived four-factor structure, we tested measurement invariance across subgroups within the sample by treatment condition, gender, and age, and across timepoints.

Treatment Invariance

We found evidence of scalar invariance in both the fall and the spring between the treatment and control groups (see Table B4 for model fits). This means that the latent factors across two different treatment groups measure the same constructs on an equivalent scale, and therefore we can directly compare treatment and control group students on the same TOOSEL constructs and on the same scale without bias.

Gender and Age Measurement Invariance

We found that TOOLSEL is scalar invariant at both waves across gender and age groups (Tables B5 and B6), which suggests that we can compare the differences by gender and age on the TOOSEL constructs without measurement bias based on a child's gender or age.

Invariance across Time

A series of longitudinal invariance models was tested to confirm that the change from the fall to the spring for the same constructs can be estimated (Table B7). Model fit difference between configural, metric, and scalar models suggested that the factor structure, loadings, and thresholds of the items were invariant from the fall to the spring. In other words, we found no significant difference in the item and measure functioning across timepoints, thus we can compare the fall and the spring scores on these constructs.

Difference of SEL across Gender, Age, and Time

Table 3 and Figures 2, 3, and 4 provided differences in TOOLSEL constructs by gender, age, and time. We found significant gender differences. Girls were rated

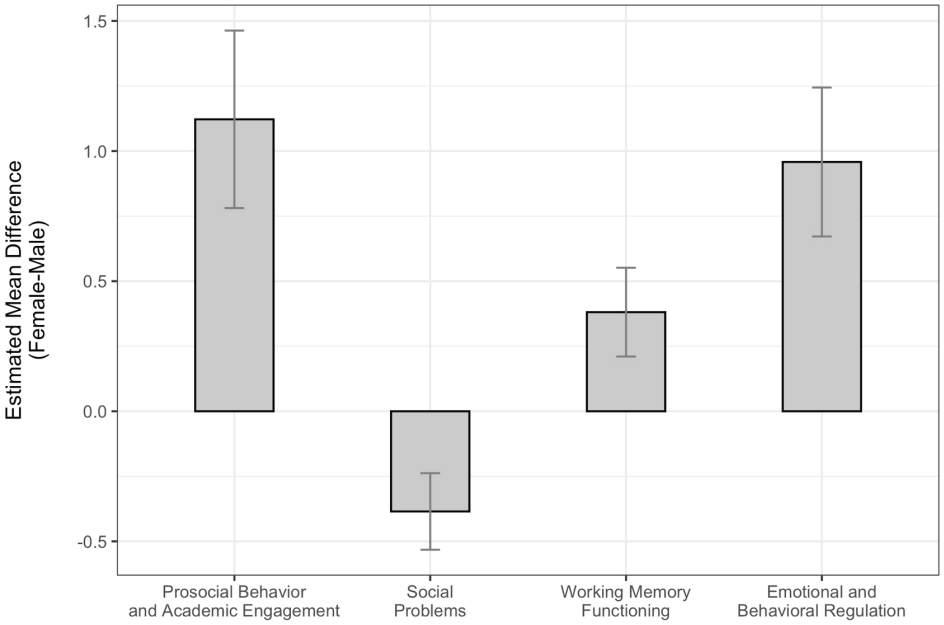
higher than boys on all the favorable TOOLSEL constructs—Prosocial Behavior and Academic Engagement, Working Memory Functioning, Emotional and Behavioral Regulation—and lower on social problems. Interestingly, we found no statistical difference by age in the TOOLSEL constructs, despite the pattern of increase in means with age. On average, teachers reported decreased Prosocial Behavior and Academic Engagement (standardized difference=-0.106, $p<.05$) and increased Social Problems (standardized difference=0.165, $p=.001$) in the spring as compared to the fall, while they did not report a significant difference in Working Memory Functioning and Emotional and Behavioral Regulation.

Table 3: Model-Based Estimates of TOOLSEL Subconstructs by Data Collection Wave, Age, and Gender

	Estimated Latent Factor Mean (SE)			
	Prosocial Behavior and Academic Engagement	Social Problems	Working Memory Functioning	Emotional and Behavioral Regulation
Data Collection				
Fall	0 (1.000)	0 (1.000)	0 (1.000)	0 (1.000)
Spring	0.106* (0.049)	-0.165** (0.055)	0.003 (0.049)	0.037 (0.053)
Age (years old)				
7 years or younger	0 (1.000)	0 (1.000)	0 (1.000)	0 (1.000)
8-9 years	0.027 (0.234)	0.108 (0.098)	0.035 (0.121)	-0.004 (0.199)
10-11 years	0.155 (0.233)	0.082 (0.106)	0.11 (0.118)	0.052 (0.185)
≥ 12 years	0.393 (0.237)	0.074 (0.11)	0.188 (0.117)	0.123 (0.197)
Gender				
Male	0 (1.000)	0 (1.000)	0 (1.000)	0 (1.000)
Female	1.122*** (0.174)	-0.385*** (0.075)	0.381*** (0.087)	0.958*** (0.146)

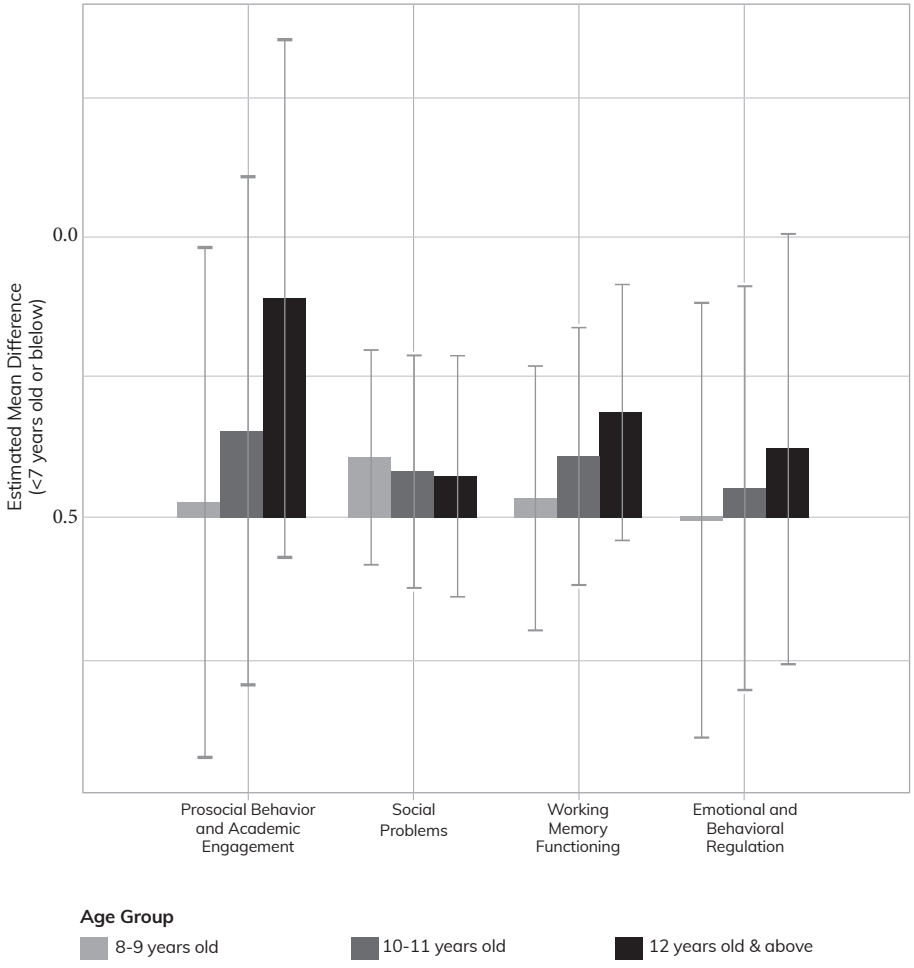
Note: In the fall, children age seven or younger and male were referenced for estimating means of other timepoints and subgroups in the models, and therefore fixed at a mean of 0 and variance of 1.
*$p<.05$, **$p<.01$, ***$p<.001$

Figure 2: Gender Differences in TOOLSEL Constructs: (1) Prosocial Behavior and Academic Engagement, (2) Social Problems, (3) Working Memory Functioning, and (4) Emotional and Behavioral Regulation



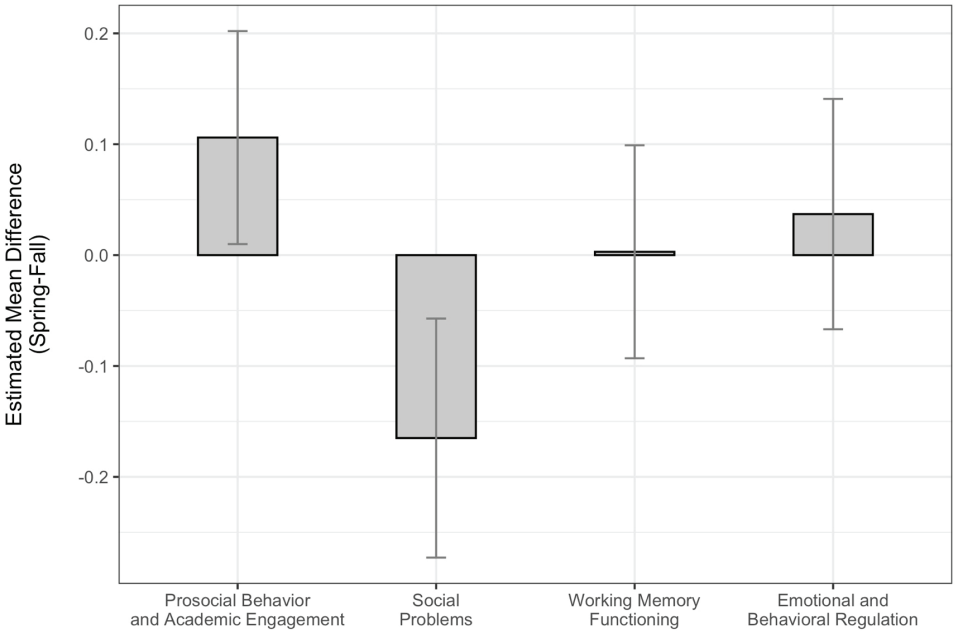
Note: Male is the reference group in estimating the mean difference.

Figure 3: Age Differences in TOOLSEL Constructs: (1) Prosocial Behavior and Academic Engagement, (2) Social Problems, (3) Working Memory Functioning, and (4) Emotional and Behavioral Regulation



Note: Children aged seven or younger is the reference group in estimating mean difference.

Figure 4: Spring-Fall Differences in TOOLSEL Constructs: (1) Prosocial Behavior and Academic Engagement, (2) Social Problems, (3) Working Memory Functioning, and (4) Emotional and Behavioral Regulation



Note: Fall is the reference group in estimating mean difference.

Correlational Evidence of Validity

Bivariate associations across time: Fall to spring

We expect teachers' perceptions of their children in a specific dimension to change somewhat, but generally to remain stable over the course of a school year. Bivariate correlations of the factor scores of all four of the TOOLSEL constructs were positively correlated across time points, r=0.585 for Prosocial Behavior and Academic Engagement, r=0.603 for Social Problems, r=0.569 for Working Memory Functioning, r=0.510 for Emotional and Behavioral Regulation. This indicates that teachers' perceptions of children's behavior were fairly consistent, displaying some continuity and some change across the six-month period (Table 4).

Table 4: Bivariate Correlations among TOOLSEL Factor Scores at Fall and Spring

	1	2	3	4	5	6	7
1. Prosocial Behavior and Academic Engagement T1	--						
2. Social Problems T1	-0.637	--					
3. Working Memory Functioning T1	0.905	-0.534	--				
4. Emotional and Behavioral Regulation T1	0.862	-0.640	0.911	--			
5. Prosocial Behavior and Academic Engagement T2	0.585	-0.456	0.527	0.487	--		
6. Social Problems T2	-0.364	0.603	-0.284	-0.376	-0.570	--	
7. Working Memory Functioning T2	0.572	-0.368	0.569	0.489	0.931	-0.427	--
8. Emotional and Behavioral Regulation T2	0.534	-0.449	0.501	0.510	0.882	-0.585	0.920

Note: All correlation coefficients were statistically significant at $p<.001$.

Bivariate Associations with Other Measures

Bivariate correlations between the TOOLSEL constructs and other external measures (Table 6) showed additional support for validity. That is, the TOOSEL constructs were correlated in the expected directions with external measures of similar constructs. The Prosocial Behavior and Academic Engagement factor was positively correlated with both the assessor report of behavioral regulation and the performance-based assessment of working memory ($r=0.147$, $p<.001$, and $r=0.152$, $p<.001$, respectively). In addition, it was negatively correlated with child self-reports of public school victimization ($r=-0.117$, $p<.001$), but not correlated with RACER inhibitory control ($r=-0.008$, $p>.05$). Social problems were positively correlated with child self-report of public school victimization ($r=0.144$, $p<.001$), as expected. However, it had a statistically significant but very small correlation with the assessor report of behavioral regulation ($r=-0.061$, $p<.001$), RACER working memory ($r=-0.050$, $p<.05$), and RACER inhibitory control ($r=0.053$, $p<.05$). TOOLSEL's Working Memory Functioning was positively correlated with the assessor report of behavioral regulation ($r=0.152$, $p<.001$) and RACER working memory ($r=0.167$, $p<.001$). In addition, Working Memory

Functioning was negatively correlated to a small degree (r=-0.091, p<.001) with child self-reported public school victimization and not correlated with RACER inhibitory control (r=0.025, p>.05). Emotional and Behavioral Regulation was positively correlated with assessor-report behavioral regulation (r=0.112, p<.001) and RACER working memory (r=0.114, p<.001), and negatively correlated with child self-report of school victimization (r=-0.128, p<.001). Interestingly, Emotional and Behavioral Regulation were not associated with the RACER inhibitory control.

Table 5: Bivariate Correlations between TOOLSEL Factor Scores and PSRA, RACER, and Victimization Scale in the Fall

	1	2	3	4	5	6	7
1. Prosocial Behavior and Academic Engagement T1	--						
2. Social Problems T1	-0.637***	--					
3. Working Memory Functioning T1	0.905***	-0.534***	--				
4. Emotional and Behavioral Regulation T1	0.862***	-0.640***	0.911***	--			
5. Public School Victimization	-0.117***	0.144***	-0.091***	-0.128***	--		
6. Behavioral Regulation	0.147***	-0.061***	0.152***	0.112***	-0.065***	--	
7. RACER Working Memory	0.152***	-0.050*	0.167***	0.114***	0.025	0.256***	--
8. RACER Inhibitory Control	-0.008	0.053*	0.025	-0.022	-0.050*	0.048	0.130***

Note: ***p<0.001, **p<0.01, *p<0.05

Partial Correlation

Table 6 presents the ordinary least squares regression models testing partial correlations between TOOLSEL constructs and other related constructs, controlling for child demographic characteristics (age, grade, gender). In addition to child demographic characteristics, measures of school victimization, working memory, inhibitory control, and behavioral regulation explained 9 percent to 12 percent of the variance in TOOLSEL constructs. Controlling for child characteristics and other measures, public school victimization was significantly associated with all TOOLSEL constructs. Specifically, a higher degree of victimization was related to lower Prosocial Behavior and Academic Engagement (b=-0.156, p<.001), lower Working Memory Functioning (b=-0.124, p<.001), lower Emotional and Behavioral Regulation (b=-0.171, p<.001), and more Social Problems (b=0.180, p<.001). Assessor-report behavioral regulation was positively related to Prosocial Behavior and Academic Engagement (b=0.104, p<.01), Working Memory Functioning (b=0.112, p<.001), and Emotional and Behavioral Regulation (b=0.090, p<.01). RACER working memory was positively associated with teacher-reported Prosocial Behavior and Academic Engagement (b=0.222, p<.001), Working Memory Functioning (b=0.234, p<.001), Emotional and Behavioral Regulation (b=0.185, p<.001), and negatively associated with Social Problems (b=-0.093, p<.01). The RACER cognitive inhibitory control measure was not related to any of the TOOSEL constructs.

Table 6: Ordinary Least Squares Regression Models Predicting TOOLSEL Constructs

	Prosocial Behavior and Academic Engagement			Social Problems			Working Memory Functioning			Emotional and Behavioral Regulation		
	Beta	*b*	*SE*	*Beta*	*b*	*SE*	*Beta*	*b*	*SE*	*Beta*	*b*	*SE*
(Intercept)	0.000	-0.199	0.167	0.000	0.073	0.169	0.000	-0.174	0.173	0.000	-0.185	0.176
Public School Victimization	-0.123	-0.156***	0.031	0.149	0.180***	0.030	-0.097	-0.124***	0.032	-0.130	-0.171***	0.032
Behavioral Regulation	0.086	0.104**	0.035	-0.017	-0.019	0.033	0.092	0.112***	0.034	0.071	0.090**	0.033
RACER Working Memory	0.137	0.222***	0.045	-0.060	-0.093*	0.041	0.143	0.234***	0.047	0.110	0.185***	0.046
RACER Inhibitory Control	-0.018	-0.017	0.022	0.046	0.041	0.023	0.010	0.009	0.023	-0.021	-0.020	0.025
Age	-0.001	0.000	0.019	0.032	0.013	0.021	-0.007	-0.003	0.020	0.006	0.003	0.020
Grade	-0.002	-0.001	0.029	-0.003	-0.002	0.032	0.009	0.006	0.033	-0.036	-0.022	0.032
Female (reference=Male)	0.257	0.498***	0.053	-0.240	-0.442***	0.056	0.223	0.437***	0.052	0.257	0.517***	0.056
R^2	0.122			0.094			0.104			0.110		

DISCUSSION

TOOLSEL was assembled from parts of existing measures to assess teachers' perceptions of students' classroom behaviors that reflect a set of social, emotional, and cognitive skills. TOOLSEL was intended to be used to evaluate a classroom-based SEL intervention for Syrian refugee children in nonformal education settings in Lebanon. Measures used to evaluate programs must meet a high standard of evidence for validity and reliability, given that the results often are used for accountability purposes and for program and policy decisionmaking that can have widespread consequences. Evidence indicates that TOOLSEL holds promise for use as a program-evaluation measure; however, we make several recommendations that would strengthen the data resulting from the use of this tool.

First, we found evidence of TOOLSEL's internal coherence, with a consistent factor structure that is meaningful and unique to the population and context. While the empirical data did not support the originally hypothesized factors for the five discrete subscales assembled across different tools, a series of exploratory and confirmatory factor analyses provided consistent support for a four-factor structure measuring teachers' perceptions of student behaviors in a classroom context: (1) Prosocial Behavior and Academic Engagement, (2) Social Problems, (3) Working Memory Functioning, and (4) Emotional and Behavioral Regulation. It is important to note that some of these final TOOLSEL subconstructs consist of items from across multiple, theoretically distinct subdomains of social and emotional skills. These results suggest that teachers are identifying the behaviors of "good" or "well-functioning" students, but not distinguishing between specific behavior subdomains; for example, prosocial versus classroom engagement behaviors (e.g., "Showing empathy" vs. "Working hard"); and emotional versus behavioral regulation skills (e.g., "Can calm down when excited or all wound up" vs. "Waits to be called on before responding"). In addition, the Prosocial Behavior and Academic Engagement subscale was highly correlated with the Working Memory Functioning and Emotional and Behavioral Regulation subscales. These findings may indicate cultural and contextual specificity in teachers' perceptions of children's social and emotional competence, and the subscales generated from this study may capture the children's skills that are better aligned with the cultural and contextual understanding of child development. On the other hand, it also may point to a limitation of teachers' reporting SEL skills. The patterns of high correlation among teacher-reported measures of related constructs are also observed in the non-EiE settings, such as the previous studies conducted in the US and Greece (Koth et al. 2009; Kourkounasiou and Skordilis 2014). Teachers are not typically trained in observing specific, distinct, social and emotional skills,

and they may rely on their global perceptions of individual children as good or bad, or as well-behaved or disruptive. This lack of specificity in teacher ratings may be important to consider when using teacher-reported measures for purposes that require an assessment of specific social, emotional, and cognitive processes.

Second, all of the empirically derived subscales for these four factors were consistent internally and over time with this sample of Syrian refugee children who were attending Lebanese public schools and taught by Lebanese teachers, which provides strong evidence of reliability. Such evidence of reliability is an important criterion for measures used for program-evaluation purposes, given that measurement error can attenuate the detection of treatment effects (Raudenbush and Sadoff 2008). Specifically, the subscales showed high internal consistency, which indicates that teachers generally provided consistent ratings on items within a subscale.

Third, we found evidence of measurement invariance with TOOLSEL by treatment, age, and gender groups, and across time (fall and spring). This means that the measure functions in the same way and is not biased against any subgroup by treatment condition, gender, or age when comparing the differences in TOOLSEL constructs. TOOLSEL also can be used without bias for program-evaluation purposes with pre- and posttest design, due to the differential functioning of the measure before and after the program implementation. In this case, some of the TOOSEL constructs showed increases (Social Problems) or decreases (Prosocial Behavior and Academic Engagement) over the duration of the program period (six months, from fall to spring). While we do not have enough information on the normative developmental patterns and change in teachers' perceptions over time for Syrian refugee children in Lebanon to determine whether these changes are in the expected direction or at the expected magnitude, these results provide some support for their use in program evaluation to detect change over the program implementation period.

Fourth, the correlational evidence provides initial support for the validity of TOOLSEL. Specifically, the four constructs showed moderate autocorrelations over the course of six months and suggested that the teachers' perceptions of children's SEL skills display some degree of continuity and some degree of change (i.e., they are relatively stable over time). While these correlations are not very high, they are aligned with US research suggesting that SEL constructs tend to be more influenced by contextual factors and are likely to vary over time, as compared to academic skills, which tend to be highly stable over time (Soland et al. 2019). We also found significant gender differences in the expected directions,

given the current literature (Zimmermann and Iwanski 2014), which suggests that TOOLSEL is sensitive to detecting teachers' perceptions of gender difference in children's SEL skills (Elzein and Ammar 2010; Keresteš 2006; Lumley et al. 2002). Specifically, teachers rated girls higher than boys on Prosocial Behavior and Academic Engagement, Working Memory Functioning, and Emotional and Behavioral Regulation factors, and lower on Social Problems. However, it was not sensitive to detecting age differences, and there is not yet evidence that TOOLSEL can be used to detect developmental differences in the SEL constructs it has been designed to measure.

In addition, teacher ratings for each of the TOOLSEL subconstructs were generally correlated with other similar concepts in the expected directions, albeit at a relatively small magnitude ($rs<0.2$). This includes an assessor-report measure of behavioral regulation, a performance-based tablet assessment of cognitive function, and child self-reports of experiencing victimization at school. It is not uncommon for reports from different raters to provide discrepant information (Buckley and Krachman 2016). While such discrepancies are often treated as a nuisance, recent research has demonstrated that discrepancies across informants can contain useful information that is helpful in interpreting program impacts, and for predicting longer-term adjustment and wellbeing (De Los Reyes 2011). While teacher reports provide meaningful information about the teachers' perception and interpretation of children's classroom behaviors, the use of multiple measurement methods and informants will be valuable in understanding children's social and emotional development in emergency contexts—especially when the purpose of assessment demands understanding children's behaviors, attitudes, and skills across multiple settings.

IMPLICATIONS FOR USE

Feasibility Considerations

Given the resource constraints common in EiE contexts, it is important to consider the field feasibility of a measure and to use caution in interpreting the evidence from teacher reports in EiE settings, for the following reasons: (1) teachers may not know students very well if the student population they serve is highly mobile or attends lessons infrequently; (2) teachers may not have time to provide thoughtful and reliable information on individual children, as they are balancing a number of competing demands—including coping with their own experiences of trauma and adversity—and also may have limited training and experience in observing

and working with children; (3) reports from teachers in refugee contexts who come from a host community with a different cultural background and context than that of the refugee children may project systematic bias against the refugees that reflects the tension between the refugee and host communities. Given these considerations, we provide several more regarding the adaptation and use of TOOLSEL.

Adaptations and Considerations for Use

While the evidence provided in this study largely supports the use of TOOLSEL for evaluation purposes with Syrian refugee children living in Lebanon, the findings are not assumptively generalizable to different populations and contexts. Hence, we strongly recommend piloting, adapting, and reevaluating the psychometric properties of the measure before using it with different populations and in different contexts. We provide a few suggestions for researchers and practitioners considering the use of TOOLSEL.

Most importantly, researchers and practitioners should ensure that the setting and structure of the program are suitable for using TOOLSEL, and that they are using it to evaluate the program's impact. TOOLSEL is designed for use in classrooms and learning spaces by teachers or facilitators who have regular and extensive interactions with individual children. This means TOOLSEL is appropriate to use with small to medium-size classes or learning groups where the children are engaged in learning activities facilitated by adults. It only can be used after the program has been launched and the teachers have had time to get to know the children well. This may not be the case in many EiE settings, where teachers often work with large groups of children and are too overwhelmed by multiple demands to get to know the children individually; moreover, children may not attend the program regularly, due to the safety and economic concerns common in EiE settings. Finally, while it may be tempting to use a measure like TOOLSEL for multiple purposes in resource-strained EiE settings, we emphasize that TOOLSEL should not be used for purposes other than program evaluation and research. Given the limited specificity of the teacher ratings we found in this report, we strongly recommend against using TOOLSEL for screening or formative assessment purposes.

Once TOOLSEL is deemed appropriate for a particular setting and purpose, we recommend a set of strategies to ensure that teachers can differentiate meaningfully between children and report on their individual behaviors in class, and thus improve the validity of the teacher-reporting scales. First, cognitive

interviewing techniques can be used during the measure pilot to understand how teachers are interpreting and responding to items, and their perceptions of the utility, reliability, and cultural and ecological validity in crisis contexts. This information can be used to refine items and assessment directions/procedures to help teachers distinguish clearly between social skills and learning-based cognitive processes, and to improve the measure's utility and validity in reflecting teachers' perspectives.

Second, explicit assessor training for teachers in filling out the survey can improve the validity of their reports. Teachers in EiE settings may not have enough experience or training to observe carefully and report on the children's individual behaviors. They also may lack sufficient literacy to understand the questions fully, especially when the written instructional language is not their first language (Dryden-Peterson 2015).[3] Therefore, establishing common understanding of the meaning of items presented in TOOLSEL for the concepts each item is intended to capture may increase the specificity of the concepts TOOLSEL can capture, and improve its reliability and validity.

Third, in planning for the assessment, we recommend implementing strategies that reduce the burden of reporting for teachers. This may include selecting a random subset of children for teachers to report on or providing coverage in the classroom to give the teacher time to fill out the survey. Fourth, we recommend using behavioral "nudge" strategies during the assessment that prime teachers to think about the many different behaviors of the focal child. Trained enumerators or tablet algorithms also could be used to quickly identify when teachers are providing a child with the same score on all items, which will result in statistics with low reliability. Fifth, we recommend that the items on the measure be adapted for each age group (i.e., early childhood, middle childhood, adolescence) so that each item is situated within an appropriate developmental trajectory. This may partially remedy the teacher reference bias and provide teachers with different forms of the measure that are based on the age of the child, rather than receiving the same measure regardless of the child's characteristics. Finally, we recommend collecting data from multiple sources to triangulate the data most effectively.

3 All teachers in our study had sufficient literacy, as their native/first language was Arabic (the language of instruction and research for this study) and they had a high school education or higher.

CONCLUSION

This study provides evidence that TOOLSEL offers coherent, reliable, consistent, and empirically valid information that is unbiased across treatment groups, gender and age groups, and the timing of the assessment. In addition, we find additional support for using TOOLSEL in program evaluation, given its ability to detect changes during a six-month implementation of the program with Syrian refugee children living in Lebanon. While testing the sensitivity to treatment is beyond the scope of this study and only can be done as a part of an impact evaluation of a program that is proven to show changes in these SEL skills, the evidence produced in the current study provides some confidence in the decision to use TOOLSEL for evaluation purposes. We acknowledge that the TOOLSEL construction relies on knowledge and tools that are based mainly on research in non-EIE contexts and thus that make a limited contribution to the decolonization of research and knowledge (Bermúdez, Muruthi, and Jordan 2016; Zavala 2013). When possible, it is more desirable to develop and adapt SEL measures that fully reflect the local context and culture and to use methodological approaches that are rooted in participant-informed coconstruction of knowledge, such as participatory research methods (Javdani, Singh, and Sichel 2017). When the tools, time, and resources needed to generate such measures are not available, TOOLSEL provides a feasible and practical alternative for assessing SEL skills that is suitable for program evaluation in EiE settings.

Indeed, research that, like this study, empirically evaluates tools or hypotheses that are developed primarily in non-EiE settings holds promise as a starting point for valuable culturally and contextually grounded research. Not all research can be built from the ground up, especially in conflict- and crisis-affected and resource-poor contexts, where the effective and prompt provision of services that support the population's urgent needs is prioritized. In such cases, this type of research can provide a practical alternative that takes the current status quo—which relies on imposing "evidence-based" knowledge from the non-EIE context—a step further toward building contextually and culturally relevant knowledge in situations and with populations that have traditionally been underrepresented, misrepresented, and marginalized.

REFERENCES

AERA (American Educational Research Association), APA (American Psychological Association), and NCME (National Council on Measurement in Education). 2014. *Standards for Educational and Psychological Testing,* 3rd ed. Washington, DC: AERA.

Asparouhov, Tihomir, and Bengt Muthén. 2010. "Weighted Least Squares Estimation with Missing Data." *Mplus Technical Appendix* (2010):1-10.

Bakrania, Shivit, Nikola Balvin, Silvio Daidone, and Jacobus de Hoop. 2021. *Impact Evaluation in Settings of Fragility and Humanitarian Emergency.* Innocenti Discussion Papers 2021-02. Florence: UNICEF Office of Research–Innocenti. https://www.unicef-irc.org/publications/pdf/Impact-Evaluation-in-Settings-of-Fragility-and-Humanitarian-Emergency.pdf.

Bermúdez, J. Maria, Bertranna A. Muruthi, and Lorien S. Jordan. 2016. "Decolonizing Research Methods for Family Science: Creating Space at the Center." *Journal of Family Theory & Review* 8 (2): 192-206. https://doi.org/10.1111/jftr.12139.

Betancourt, Theresa, Sarah Meyers-Ohki, Alexandra Charrow, and Wietse Tol. 2013. "Interventions for Children Affected by War: An Ecological Perspective on Psychosocial Support and Mental Health Care." *Harvard Review of Psychiatry* 21 (2): 70-91. https://doi.org/10.1097/HRP.0b013e318283bf8f.

Blair, Clancy, and Rachel Razza. 2007. "Relating Effortful Control, Executive Function, and False Belief Understanding to Emerging Math and Literacy Ability in Kindergarten." *Child Development* 78 (2): 647-63. https://doi.org/10.1111/j.1467-8624.2007.01019.x.

Boekaerts, Monique, and Reinhard Pekrun. 2015. "Emotions and Emotion Regulation in Academic Settings." In *Handbook of Educational Psychology,* edited by Lyn Corno and Eric M. Anderman, 90-104. New York: Routledge.

Brody, Leslie. 2000. "The Socialization of Gender Differences in Emotional Expression: Display Rules, Infant Temperament, and Differentiation." *Gender and Emotion: Social Psychological Perspectives* 2: 24-47. https://doi.org/10.1017/CBO9780511628191.003.

Buckley, Katie, and Sara Krachman. 2016. *Initial Findings from the Boston Charter Research Collaborative*. Boston, MA: Transforming Education. https://www.transformingeducation.org/wp-content/uploads/2017/04/TE-BCRCWorkingPaperFINAL.pdf.

Burde, Dana, Ozen Guven, Jo Kelcey, Heddy Lahmann, and Khaled Al-Abbadi. 2015. "What Works to Promote Children's Educational Access, Quality of Learning, and Wellbeing in Crisis-Affected Contexts." Education Literature Review. London: Department for International Development. https://assets.publishing.service.gov.uk/media/57a0897ee5274a31e00000e0/61127-Education-in-Emergencies-Rigorous-Review_FINAL_2015_10_26.pdf.

Burde, Dana, Amy Kapit, Rachel Wahl, Ozen Guven, and Margot Skarpeteig. 2017. "Education in Emergencies: A Review of Theory and Research." *Review of Educational Research* 87 (3): 619-58. https://doi.org/10.3102/0034654316671594.

Catalán, Héctor. 2019. "Reliability, Population Classification and Weighting in Multidimensional Poverty Measurement: A Monte Carlo Study." *Social Indicators Research* 142 (3): 887-910. https://doi.org/10.1007/s11205-018-1950-z.

Chen, Fang Fang. 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.

Cole, Pamela, Margaret Michel, and Laureen O'Donnell Teti. 1994. "The Development of Emotion Regulation and Dysregulation: A Clinical Perspective." *Monographs of the Society for Research in Child Development* 59 (2-3): 73-102.

Conduct Problems Prevention Research Group. 1990. "Social Competence Scale (Teacher Version)." *Zugriff Am* 5: 2007.

Davies, Lynn, and Christopher Talbot. 2008. "Learning in Conflict and Postconflict Contexts." *Comparative Education Review* 52 (4): 509-18. https://doi.org/10.1086/591295.

De Los Reyes, Andres. 2011. "Introduction to the Special Section: More Than Measurement Error. Discovering Meaning behind Informant Discrepancies in Clinical Assessments of Children and Adolescents." *Journal of Clinical Child & Adolescent Psychology* 40 (1): 1-9. https://doi.org/10.1080/15374416.2011.533405.

De Los Reyes, Andres, Tara Augenstein, Mo Wang, Sarah Thomas, Deborah Drabick, Darcy Burgers, and Jill Rabinowitz. 2015. "The Validity of the Multi-Informant Approach to Assessing Child and Adolescent Mental Health." *Psychological Bulletin* 141 (4): 858. https://doi.org/10.1037/a0038498.

Duncan, Robert, Megan McClelland, and Alan Acock. 2017. "Relations between Executive Function, Behavioral Regulation, and Achievement: Moderation by Family Income." *Journal of Applied Developmental Psychology* 49: 21-30. https://doi.org/10.1016/j.appdev.2017.01.004.

Durlak, Joseph, Roger Weissberg, Allison Dymnicki, Rebecca Taylor, and Kriston Schellinger. 2011. "The Impact of Enhancing Students' Social and Emotional Learning: A Meta-Analysis of School-based Universal Interventions." *Child Development* 82 (1): 405-32. https://doi.org/10.1111/j.1467-8624.2010.01564.x.

Eisenberg, Nancy, Cynthia L. Smith, and Tracy L. Spinrad. 2011. "Effortful Control: Relations with Emotion Regulation, Adjustment, and Socialization in Childhood." In *Handbook of Self-Regulation: Research, Theory, and Applications,* 2nd ed., edited by Kathleen D. Vohs and Roy F. Baumeister, 263-83. New York: Guilford Press.

Ellis, Bruce, Anthony Volk, Jose-Michael Gonzalez, and Dennis Embry. 2016. "The Meaningful Roles Intervention: An Evolutionary Approach to Reducing Bullying and Increasing Prosocial Behavior." *Journal of Research on Adolescence* 26 (4): 622-37. https://doi.org/10.1111/jora.12243.

Elzein, Heyam, and Diala Ammar. 2010. "Parent and Teacher Perceptions of Assessing Lebanese Children's Reaction to War-Related Stress: A Survey of Psychological and Behavioral Functioning." *Journal of Child & Adolescent Trauma* 3 (4): 255-78. https://doi.org/10.1080/19361521.2010.523060.

Espelage, Dorothy, and Melissa Holt. 2001. "Bullying and Victimization during Early Adolescence: Peer Influences and Psychosocial Correlates." *Journal of Emotional Abuse* 2 (2-3): 123-42. https://doi.org/10.1300/J135v02n02_08.

Fabes, Richard, and Nancy Eisenberg. 1998. "Meta-Analyses of Age and Sex Differences in Children's and Adolescents' Prosocial Behavior." *Handbook of Child Psychology* 3: 1-29.

Ford, Cassie, Ha Yeon Kim, Lindsay Brown, J. Lawrence Aber, and Margaret Sheridan. 2019. "A Cognitive Assessment Tool Designed for Data Collection in the Field in Low- and Middle-Income Countries." *Research in Comparative and International Education* 14 (1): 141-57. https://doi.org/10.1177/1745499919829217.

Furrer, Carrie, and Ellen Skinner. 2003. "Sense of Relatedness as a Factor in Children's Academic Engagement and Performance." *Journal of Educational Psychology* 95 (1): 148-62. https://doi.org/10.1037/0022-0663.95.1.148.

Gest, Scott D., Janet A. Welsh, and Celene E. Domitrovich. 2005. "Behavioral Predictors of Changes in Social Relatedness and Liking School in Elementary School." *Journal of School Psychology* 43 (4): 281-301. https://doi.org/10.1016/j.jsp.2005.06.002.

Goldman-Rakic, Patricia. 1996. "Regional and Cellular Fractionation of Working Memory." *Proceedings of the National Academy of Sciences* 93 (24): 13473-80. https://doi.org/10.1073/pnas.93.24.13473.

Hamoudi, Amar, and Margaret Sheridan. 2015. "Unpacking the Black Box of Cognitive Ability: A Novel Tool for Assessment in a Population-Based Survey." Manuscript under review, Young Lives Study.

Hartup, Willard. 1996. "The Company They Keep: Friendships and Their Developmental Significance." *Child Development* 67 (1): 1-13. https://doi.org/10.1111/j.1467-8624.1996.tb01714.x.

Hayes, Andrew, and Jacob Coutts. 2020. "Use Omega Rather than Cronbach's Alpha for Estimating Reliability, But…" *Communication Methods and Measures* 14 (1): 1-24. https://doi.org/10.1080/19312458.2020.1718629.

Heckman, James, Jora Stixrud, and Sergio Urzua. 2006. "The Effects of Cognitive and Noncognitive Abilities on Labor Market Outcomes and Social Behavior." *Journal of Labor Economics* 24 (3): 411-82. https://doi.org/10.1086/504455.

Hu, Li-tze, and Peter Bentler. 1999. "Cutoff Criteria for Fit Indexes in Covariance Structure Analysis: Conventional Criteria versus New Alternatives." *Structural Equation Modeling: A Multidisciplinary Journal* 6 (1): 1-55. https://doi.org/10.1080/10705519909540118.

Hughes, Claire, and Rosie Ensor. 2011. "Individual Differences in Growth in Executive Function across the Transition to School Predict Externalizing and Internalizing Behaviors and Self-Perceived Academic Success at 6 Years of Age." *Journal of Experimental Child Psychology* 108 (3): 663-76. https://doi.org/10.1016/j.jecp.2010.06.005.

Jacob, Robin, and Julia Parkinson. 2015. "The Potential for School-Based Interventions That Target Executive Function to Improve Academic Achievement: A Review." *Review of Educational Research* 85 (4): 512-52. https://doi.org/10.3102/0034654314561338.

Javdani, Shabnam, Sukhmani Singh, and Corianna Sichel. 2017. "Negotiating Ethical Paradoxes in Conducting a Randomized Controlled Trial: Aligning Intervention Science with Participatory Values." *American Journal of Community Psychology* 60 (3-4): 439-49. https://doi.org/10.1002/ajcp.12185.

Jones, Damon, Mark Greenberg, and Max Crowley. 2015. "Early Social-Emotional Functioning and Public Health: The Relationship Between Kindergarten Social Competence and Future Wellness." *American Journal of Public Health* 105 (11): 2283-90. https://doi.org/10.2105/AJPH.2015.302630.

Jones, Stephanie, Rebecca Bailey, and Sophie Barnes. 2015. *Classroom Executive Function Survey (CEFS): A Measure of Executive Function and Self-Regulation Skills in Everyday Classroom Behavior.* Cambridge, MA: Harvard University Press.

Jordans, Mark, Hugo Pigott, and Wietse Tol. 2016. "Interventions for Children Affected by Armed Conflict: A Systematic Review of Mental Health and Psychosocial Support in Low- and Middle-Income Countries." *Current Psychiatry Reports* 18 (1): 1-15. https://doi.org/10.1007/s11920-015-0648-z.

Keresteš, Gordana. 2006. "Children's Aggressive and Prosocial Behavior in Relation to War Exposure: Testing the Role of Perceived Parenting and Child's Gender." *International Journal of Behavioral Development* 30 (3): 227-39. https://doi.org/10.1177/0165025406066756.

Kim, Ha Yeon, Lindsay Brown, Carly Tubbs Dolan, Kalina Gjicali, Rena Deitz, Sol Prieto Bayona, and J. Lawrence Aber. "Comprehensive Social Emotional Learning Intervention with Syrian Refugee Children: Impact Variation by Pre- and Post-Migration Conflict Experiences." Unpublished manuscript, 2021.

Margaret Sheridan, and John Lawrence Aber. 2020. "Post-Migration Risks, Developmental Processes, and Learning among Syrian Refugee Children in Lebanon." *Journal of Applied Developmental Psychology* 69: 101142. https://doi.org/10.1016/j.appdev.2020.101142.

Kochanska, Grazyna, Kathleen Murray, and Elena Harlan. 2000. "Effortful Control in Early Childhood: Continuity and Change, Antecedents, and Implications for Social Development." *Developmental Psychology* 36 (2): 220-32. https://doi.org/10.1037/0012-1649.36.2.220.

Koth, Christine W., Catherine P. Bradshaw, and Philip J. Leaf. 2009. "Teacher Observation of Classroom Adaptation—Checklist: Development and Factor Structure." *Measurement and Evaluation in Counseling and Development* 42 (1): 15-30. https://doi.org/10.1177/0748175609333560.

Kourkounasiou, Mari, and Emmanouil Skordilis. 2014. "Validity and Reliability Evidence of the TOCA-C in a Sample of Greek Students." *Psychological Reports* 115 (3): 766-83. https://doi.org/10.2466/08.11.PR0.115c31z5.

LaFontana, Kathryn, and Antonius Cillessen. 2002. "Children's Perceptions of Popular and Unpopular Peers: A Multimethod Assessment." *Developmental Psychology* 38 (5): 635-47. https://doi.org/10.1037/0012-1649.38.5.635.

Lei, Pui-Wa. 2009. "Evaluating Estimation Methods for Ordinal Data in Structural Equation Modeling." *Quality and Quantity* 43 (3): 495. https://doi.org/10.1007/s11135-007-9133-z.

Lumley, Vicki, Cheryl McNeil, Amy Herschell, and Alisa Bahl. 2002. "An Examination of Gender Differences among Young Children with Disruptive Behavior Disorders." *Child Study Journal* 32 (2): 89-101.

McCoy, Dana Charles, Stephanie Zuilkowski, Hirokazu Yoshikawa, and Günther Fink. 2017. "Early Childhood Care and Education and School Readiness in Zambia." *Journal of Research on Educational Effectiveness* 10 (3): 482-506. https://doi.org/10.1080/19345747.2016.1250850.

McDonald, Roderick P. 1999. *Test Theory: A Unified Treatment*. New Jersey: Lawrence Erlbaum Associates.

McKown, Clark, and Rhona Weinstein. 2008. "Teacher Expectations, Classroom Context, and the Achievement Gap." *Journal of School Psychology* 46 (3): 235-61. https://doi.org/10.1016/j.jsp.2007.05.001.

McRae, Kateri, Kevin Ochsner, Iris Mauss, John Gabrieli, and James Gross. 2008. "Gender Differences in Emotion Regulation: An FMRI Study of Cognitive Reappraisal." *Group Processes & Intergroup Relations* 11 (2): 143-62. https://doi.org/10.1177/1368430207088035.

Muthén, Bengt, and Linda Muthén. 2014. MPlus (version 7.2). Computer software. Los Angeles: Muthén and Muthén.

Raudenbush, Stephen, and Sally Sadoff. 2008. "Statistical Inference When Classroom Quality Is Measured with Error." *Journal of Research on Educational Effectiveness* 1 (2): 138-54. https://doi.org/10.1080/19345740801982104.

Raver, Cybele, Stephanie Jones, Christine Li-Grining, Fuhua Zhai, Kristen Bub, and Emily Pressler. 2011. "CSRP's Impact on Low-Income Preschoolers' Preacademic Skills: Self-Regulation as a Mediating Mechanism." *Child Development* 82 (1): 362-78. https://doi.org/10.1111/j.1467-8624.2010.01561.x.

Reise, Steven P., Keith F. Widaman, and Robin H. Pugh. 1993. "Confirmatory Factor Analysis and Item Response Theory: Two Approaches for Exploring Measurement Invariance." *Psychological Bulletin* 114 (3): 552.

Revelle, William, and Richard Zinbarg. 2009. "Coefficients Alpha, Beta, Omega, and the Glb: Comments on Sijtsma." *Psychometrika* 74 (1): 145.

Rothbart, Mary K., and M. Rosario Rueda. 2005. "The Development of Effortful Control." In *Developing Individuality in the Human Brain: A Tribute to Michael I. Posner,* edited by Ulrich Mayr, Edward Awh, and Steven W. Keele, 167-88. Washington, DC: American Psychological Association. https://doi.org/10.1037/11108-009.

Shah, Ritesh. 2017. *Improving Children's Wellbeing: An Evaluation of NRC's Better Learning Programme in Palestine.* Oslo: Norwegian Refugee Council. https://www.nrc.no/globalassets/pdf/evaluations/nrc-blp-palestine-full-report.pdf.

Simon, Richard, and Alan Rudell. 1967. "Auditory S-R Compatibility: The Effect of an Irrelevant Cue on Information Processing." *Journal of Applied Psychology* 51 (3): 300. https://doi.org/10.1037/h0020586.

Smith-Donald, Radiah, Cybele Raver, Tiffany Hayes, and Breeze Richardson. 2007. "Preliminary Construct and Concurrent Validity of the Preschool Self-Regulation Assessment (PSRA) for Field-Based Research." *Early Childhood Research Quarterly* 22 (2): 173-87. https://doi.org/10.1016/j.ecresq.2007.01.002.

Valiente, Carlos, Nancy Eisenberg, R. G. Haugen, Tracy Spinrad, Claire Hofer, Jeffrey Liew, and Anne Kupfer. 2011. "Children's Effortful Control and Academic Achievement: Mediation through Social Functioning." *Early Education & Development* 22 (3): 411-33. https://doi.org/10.1080/10409289.2010.505259.

Van de Mortel, Thea F. 2008. "Faking It: Social Desirability Response Bias in Self-Report Research." *Australian Journal of Advanced Nursing* 25 (4): 40.

Vandenberg, Robert, and Charles Lance. 2000. "A Review and Synthesis of the Measurement Invariance Literature: Suggestions, Practices, and Recommendations for Organizational Research." *Organizational Research Methods* 3 (1): 4-70. https://doi:10.1177/109442810031002.

Zavala, Miguel. 2013. "What Do We Mean by Decolonizing Research Strategies? Lessons from Decolonizing, Indigenous Research Projects in New Zealand and Latin America." *Decolonization: Indigeneity, Education & Society* 2 (1): 55-71.

Zelazo, Philip David, Stephanie M. Carlson, and Amanda Kesek. 2008. "The Development of Executive Function in Childhood." In *Handbook of Developmental Cognitive Neuroscience,* 2nd ed., edited by Charles A. Nelson and Monica Luciana, 553-74. Cambridge, MA: MIT Press.

Zimmermann, Peter, and Alexandra Iwanski. 2014. "Emotion Regulation from Early Adolescence to Emerging Adulthood and Middle Adulthood: Age Differences, Gender Differences, and Emotion-Specific Developmental Variations." International Journal of Behavioral Development 38 (2): 182-94. https://doi.org/10.1177/0165025413515405.

APPENDIX A

TOOLSEL Measure Item Description and Descriptive Statistics

Table A1: TOOLSEL Measure Descriptions

Item	Description
1	TOC1: In the last two weeks [your child]: Concentrates
2	TOC2: In the last two weeks [your child]: Is friendly
3	TOC3: In the last two weeks [your child]: Pays attention
4	TOC4: In the last two weeks [your child]: Breaks rules (removed)
5	TOC5: In the last two weeks [your child]: Is liked by classmates
6	TOC7: In the last two weeks [your child]: Works hard
7	TOC9: In the last two weeks [your child]: Shows empathy & compassion for other's feelings
8	TOC10: In the last two weeks [your child]: Gets angry when provoked by other children
9	TOC11: In the last two weeks [your child]: Stay on task (removed)
10	TOC12: In the last two weeks [your child]: Yells at others
11	TOC14: In the last two weeks [your child]: Is rejected by classmates
12	TOC15: In the last two weeks [your child]: Fights
13	TOC17: In the last two weeks [your child]: Has many friends (removed)
14	TOC20: In the last two weeks [your child]: Teases classmates
15	TOC21: In the last two weeks [your child]: Learns up to ability
16	CEFS1: In the last two weeks [your child]: Remembers lists or items in the correct order
17	SCS2: In the last two weeks [your child]: Can accept things not going his/her way (removed)

Item	Description
18	CEFS2: In the last two weeks [your child]: Follows multiple-step instructions
19	CEFS3: In the last two weeks [your child]: Uses multiple rules to complete a task
20	SCS8: In the last two weeks [your child]: Thinks before acting (removed)
21	CEFS4: In the last two weeks [your child]: Waits to be called on before responding
22	SCS11: In the last two weeks [your child]: Can calm down when excited or all wound up
23	CEFS6: In the last two weeks [your child]: Transitions easily to new activities, tasks, or major parts of the day (e.g., from recess)
24	CEFS8: In the last two weeks [your child]: Uses self-control techniques
25	SCS12: In the last two weeks [your child]: Can wait in line patiently when necessary
26	CEFS9: In the last two weeks [your child]: Waits patiently for her/his turn
27	CEFS10: In the last two weeks [your child]: Uses listening skills
28	SCS18: In the last two weeks [your child]: Controls temper when there is a disagreement

Note: Items labeled starting with TOC are taken from TOCA-C, with original item numbers used in TOCA-C. Similarly, items labeled starting with SCS were taken from the SCS Emotional Regulation Scale, and items labeled starting with CEFS were taken from CEFS. Some items on this list were removed from the final scale, as indicated.

Table A2: Descriptive Statistics of Indicators by Proposed Construct

Item	Fall (N=3,254) N	Spring (N=2,952) M	*SD*	Min	Max	N	*M*	*SD*	Min	Max
TOC1	3254	3.632	1.103	1	5	2950	3.536	1.133	1	5
TOC2	3248	3.823	1.015	1	5	2947	3.680	1.055	1	5
TOC3	3246	3.673	1.092	1	5	2942	3.533	1.146	1	5
TOC4	3233	2.467	1.151	1	5	2942	2.359	1.120	1	5
TOC5	3223	3.764	0.966	1	5	2933	3.638	1.025	1	5
TOC7	3227	3.592	1.064	1	5	2924	3.487	1.102	1	5
TOC9	3212	3.597	1.038	1	5	2922	3.505	1.067	1	5
TOC10	3224	2.717	1.248	1	5	2929	2.823	1.207	1	5
TOC11	3204	3.419	1.129	1	5	2912	3.370	1.140	1	5
TOC12	3208	2.112	1.170	1	5	2923	2.259	1.163	1	5
TOC14	3229	1.850	1.065	1	5	2926	1.988	1.086	1	5
TOC15	3231	2.005	1.190	1	5	2940	2.158	1.194	1	5
TOC17	3215	3.485	1.098	1	5	2919	3.487	1.091	1	5
TOC20	3207	2.183	1.242	1	5	2913	2.240	1.223	1	5
TOC21	3208	3.504	1.069	1	5	2924	3.434	1.060	1	5
SCS2	3230	3.540	1.056	1	5	2935	3.430	1.053	1	5
SCS8	3232	3.468	1.094	1	5	2942	3.418	1.103	1	5
SCS11	3221	3.518	1.113	1	5	2932	3.449	1.071	1	5
SCS12	3214	3.530	1.105	1	5	2929	3.457	1.074	1	5

Item	Fall (N=3,254) N	Spring (N=2,952) M	*SD*	Min	Max	N	*M*	*SD*	Min	Max
SCS18	3242	3.535	1.185	1	5	2947	3.447	1.138	1	5
CEFS1	3218	3.534	1.050	1	5	2933	3.462	1.071	1	5
CEFS2	3241	3.609	1.092	1	5	2944	3.518	1.057	1	5
CEFS3	3235	3.348	1.127	1	5	2946	3.382	1.097	1	5
CEFS4	3234	3.521	1.105	1	5	2937	3.454	1.095	1	5
CEFS6	3244	3.587	1.073	1	5	2941	3.533	1.062	1	5
CEFS8	3234	3.443	1.097	1	5	2943	3.430	1.057	1	5
CEFS9	3240	3.535	1.122	1	5	2940	3.473	1.084	1	5
CEFS10	3239	3.606	1.099	1	5	2944	3.545	1.089	1	5

Note: Items labeled starting with TOC are taken from TOCA-C, with original item numbers used in TOCA-C. Similarly, items labeled starting with SCS were taken from SCS Emotional Regulation Scale, and items labeled starting with CEFS were taken from CEFS.

APPENDIX B

Model Fit Indices

Table B1: Model Fit Indices of Confirmatory Factor Analyses of Originally Proposed Subscales (five-factor models)

Wave	k	Chi-sq	df	p	CFI	TLI	RMSEA	SRMR
Fall	150	4068.026	340	0	0.929	0.921	0.082	0.051
Spring	150	4312.336	340	0	0.929	0.921	0.089	0.055

Table B2: Model Fit Indices of Exploratory Factor Analyses Four-Factor Models for Fall and Spring

Wave	CFI	TLI	RMSEA	SRMR
Fall	0.965	0.951	0.060	0.025
Spring	0.965	0.951	0.065	0.023

Table B3: Model Fit Indices of Confirmatory Factor Analyses Final Models for Fall and Spring

Wave	**k**	**Chi-sq**	**df**	***p***	**CFI**	**TLI**	**RMSEA**	**SRMR**
Fall	123	1529.214	222	0	0.972	0.968	0.06	0.029
Spring	123	1636.506	222	0	0.972	0.968	0.066	0.037

Table B4: Model Fit Indices of Treatment Invariance Models

Model	k	χ^2	**df**	**p**	$\Delta\chi^2$	**df**	***p***	**CFI**	**TLI**	**RMSEA**	**SRMR**
Fall											
Configural	246	2255.002	444	0				0.967	0.962	0.05	0.032
Metric	227	1711.771	463	0	40.906	19	0.0025	0.977	0.975	0.041	0.033
Scalar	139	1720.471	551	0	130.661	88	0.0022	0.979	0.98	0.036	0.034
Spring											
Configural	246	2718.794	444	0				0.96	0.955	0.059	0.038
Metric	227	2132.169	463	0	65.362	19	0	0.971	0.968	0.049	0.04
Scalar	139	2183.087	551	0	206.47	88	0	0.972	0.974	0.045	0.041

Table B5: Model Fit Indices of Gender Invariance Models

Model	k	χ^2	df	p	$\Delta\chi^2$	df	*p*	CFI	TLI	RMSEA	SRMR
Fall											
Configural	246	3170	444	0				0.961	0.956	0.061	0.033
Metric	227	2315	463	0	51.89	19	0	0.973	0.971	0.050	0.035
Scalar	139	2289	551	0	122.28	88	0.001	0.975	0.977	0.044	0.035
Spring											
Configural	246	3582	444	0				0.965	0.960	0.069	0.038
Metric	227	2727	463	0	61.03	19	0	0.975	0.973	0.058	0.039
Scalar	139	2728	551	0	185.34	88	0	0.976	0.978	0.052	0.040

Table B6: Model Fit Indices of Age Invariance Models

Model	k	χ2	df	p	Δχ2	df	p	CFI	TLI	RMSEA	SRMR
Fall											
Configural	492	3488	888	0				0.973	0.969	0.060	0.033
Metric	435	2536	945	0	107.86	57	.0001	0.983	0.982	0.046	0.035
Scalar	171	2696	1209	0	277.49	264	0.272	0.984	0.987	0.039	0.036
Spring											
Configural	492	3837	888	0				0.975	0.972	0.067	0.037
Metric	435	2872	945	0	123.15	57	0	0.984	0.983	0.053	0.039
Scalar	171	3072	1209	0	357.55	264	0	0.984	0.987	0.046	0.040

Table B7: Model Fit Indices of Longitudinal Invariance Models

Model	**k**	χ^2	**df**	**p**	$\Delta\chi^2$	**df**	**p**	**CFI**	**TLI**	**RMSEA**	**SRMR**
Configural	262	3530.362	957	0				0.966	0.963	0.028	0.03
Metric	243	3163.245	976	0	35.621	19	0.0117	0.971	0.969	0.025	0.03
Scalar	155	3292.212	1064	0	211.309	107	0	0.97	0.971	0.025	0.031