Self-Regulation Assessment-Assessor Report (SRA-AR): Psychometric Evidence from Syrian Refugee Children in Lebanon

Zezhen Wu, Kalina Gjicali, Ha Yeon Kim[1], & Carly Tubbs Dolan

NYU Global TIES for Children

Technical Working Paper

December 2020

Abstract

This study aims to provide evidence on the psychometric properties of the Self-Regulation Assessment-Assessor Report (SRA-AR), a measurement tool used to capture assessors' perceptions of Syrian refugee children's skills at regulating their behavior during an assessment. SRA-AR was adapted based on the Assessor Report in the Preschool Self-Regulation Assessment (PSRA-AR) by Smith-Donald et al. (2007). This study used data collected from a large sample of 4,598 Syrian refugee children aged 5-15 enrolled in Lebanese public schools in two governorates in Lebanon. Psychometric analyses indicated that the measure captures one uniform dimension of behavioral regulation. We provide evidence that SRA-AR measures behavioral regulation with good reliability. We also provide evidence that the measure functioned and was understood in the same way by children: with access and without access to social and emotional learning (SEL) programming; at the beginning, middle, and the end of the school year; across gender; and across ages. In addition, scores on SRA-AR were moderately correlated across the three time points within a school year, suggesting relative stability of assessor-reported children's behavioral regulation over time.

[1] Correspondence should be directed to Ha Yeon Kim at hayeon@nyu.edu
Global TIES for Children, New York University
627 Broadway, Room 807, New York, NY, USA 10012 | Steinhardt.nyu.edu/ihdsc/global-ties

Overview of SRA-AR: MENAT Measurement Library Criteria

SRA-AR should have moderate to high evidence of validity/reliability for use as an program evaluation measure. Much of the evidence is positive, and we are confident in the quality of the evidence. We recommend that future studies additionally test the inter-rater reliability of the measure and dedicate time in enumerator training to ensure enumerator understanding of the items and response options. Given the strength of the evidence, we recommend this measure for its specified purpose(s) with adaptations for context and attention to the recommendations in the report (see p. 21).

Criteria	Indicators	Notes
Purpose	Program Evaluation	Requires high internal consistency and ideally interrater reliability; strong evidence of validity; sensitivity to change; ideally measurement invariance
Empirical evidence overall	# of types of evidence available	7
	% of evidence meets empirical criteria	100%
	Evidence fit for purpose	Yes for validity, reliability, and measurement invariance. No evidence of inter-rater reliability presented.
Confidence in evidence	Sampling method	Stratified by region and randomized into one of the three treatment arms
	Sample size	Large (~ 4598)
	Missing data	Missing data addressed using rigorous methods
	Rigor of method	High
Revisions	Clear guidance on what to adjust/refine	Yes

Constructs/sub-constructs assessed	Internal structural validity	Correlational validity	Internal consistency	Measurement invariance			
				Gender	Treatment	Age	Time
Behavioral regulation skills	✓	✓	✓	✓	✓	✓	✓

Key

✓	Good/excellent evidence against empirical criteria	○	Fair/inconclusive evidence against empirical criteria	✗	Little to no evidence against empirical criteria	NA	Not applicable

For additional information on the empirical criteria, please see https://inee.org/measurement-library/measure-review-criteria.

This technical working paper was developed by Zezhen Wu, Kalina Gjicali, Ha Yeon Kim, & Carly Tubbs Dolan as part of the Education in Emergencies: Evidence for Action (3EA) initiative, a research-practice partnership with the International Rescue Committee. It was reviewed by NYU Global Ties for Children for inclusion in the INEE Measurement Library. We thank children and enumerators in Lebanon, IRC Lebanon country office staff, and IRC headquarters staff, without whom this study would not have been possible. We would also like to acknowledge the support provided by E-Cubed and an anonymous private donor to create and publish this report.

Suggested citation: Wu, Z., Gjicali, K., Kim, H.Y., & Tubbs Dolan, C. (2020, December). *Self-Regulation Assessment-Assessor Report (SRA-AR): Psychometric evidence from Syrian refugee children in Lebanon.* Technical working paper. New York, NY: New York University.

Introduction

Behavioral regulation—the degree to which students can modulate their behavior and/or emotional state towards a specific goal within day-to-day environments—is a foundational process that allows children to successfully adjust to and learn in schools (Duncan et al., 2017; Ursache et al., 2012). This "modulating system" (Smith-Donald et al., 2007) includes broader behavior management strategies such as impulse control, delay of gratification, or acts of compliance rather than defiance. Given the high-risk settings in which Syrian refugee children reside and learn, they may be particularly subject to difficulties with behavioral dysregulation and its consequences. Meanwhile, strong self-regulatory skills can protect these children against the negative impacts of adversity. For instance, studies conducted in the U.S. have found substantial evidence that low-income children with better self-regulatory skills are more resilient to developmental and psychological adversities (Blair, 2010). Studies have also found that behavioral regulation is associated with children's later academic achievement, interpersonal skills, and mental and physical health (Pandey et al., 2018; Robson et al., 2020). If we could determine that behavioral regulation plays a central role in refugee children's academic success, it would also serve as an important target for interventions that support resilient responses to adverse experiences in school.

While there is little research on how behavioral regulation operates in the Syrian refugee in Lebanon context, experiences of war, violence, conflict, and poverty likely pose significant challenges to children's ability to adjust, regulate and learn in a new environment (Khamis, 2019). Furthermore, studies have found that exposure to interpersonal and community violence—in part due to heightened attention to threat (Dodge et al., 1995) and parental stress associated with the perception of an unsafe environment (Linares et al., 2001)—is related to behavioral dysregulation. Therefore, it is crucial for researchers to provide reliable and valid measures of children's behavioral regulation in a humanitarian context to meaningfully understand and support the development of behavioral regulation skills for refugee children.

This report presents the psychometric evidence of a short version of the Self-Regulation Assessment-Assessor Report (SRA-AR), an assessor-report measure of behavioral regulation. SRA-AR consists of a subset of items taken from the post-assessment, assessor-report section of the Preschool Self-Regulation Assessment-Assessor Report (PSRA-AR: Smith-Donald et al., 2007), a performance-based measure originally designed to assess self-regulation skills of preschool children in the U.S. We use the short 13-item version adapted for a study in Zambia (McCoy et al., 2017), and present the evidence from primary school-aged Syrian refugee children enrolled in a non-formal remedial education and social and emotional learning (SEL) program in

Lebanon. We evaluate whether SRA-AR is adequate for program evaluation purposes, based on its evidence of validity, reliability, and measurement invariance across different groups, including treatment groups, gender, age, and time. Building a measure that meets such psychometric standards enables stakeholders to better understand and build confidence in program impacts, which is particularly important given that such evidence is often used for accountability purposes and for program and policy decision-making that can have widespread consequences.

Research Aims

In this report, we provide evidence on the validity and reliability of a shortened and adapted version of the SRA-AR measure used with Syrian refugee children enrolled in public schools in Lebanon (SRA-AR). Through our analysis, we aim to provide:

1. **Structural evidence of validity and reliability**, including (a) evidence of the extent to which the internal factor structure of items is consistent with the constructs that the measure was intended to assess; and (b) evidence of internal consistency.
2. **Correlational evidence of validity,** by examining correlations among the behavioral regulation scores across time; and presenting evidence of correlation with related constructs.
3. **Evidence for measurement invariance** across (a) treatment groups, (b) child gender, (c) child age groups (< 8: early childhood; 8-9: early middle childhood, 10-11: middle childhood, >12: adolescence), and (d) time (i.e., longitudinal invariance, across fall, spring, and summer).

Methods

Sample

All data presented here were collected as a part of a large-scale and multi-year cluster randomized controlled trials (cRCT) of non-formal remedial and SEL programming provided by the International Rescue Committee (IRC) to Syrian refugee children in Lebanon in school years (SY) 2016-2018. The data we draw on for this report were collected from students who participated in the evaluation study in SY2016-2017. The 87 community-based sites recruited in the Akkar (N = 43) and Bekaa (N = 44) regions were stratified by region and randomized into one of the three treatment arms: 21 waitlist control sites (WC), 33 Healing Classrooms Basic Remedial Support sites (HCR), and 33 Healing Classrooms Remedial Support + Targeted Mindfulness (HCR + Mind) sites (see Tubbs Dolan et al., revise and resubmit, for additional details on the design and results of the cRCT). The data were collected at baseline (December), midline (March), and endline (May). Children were assessed by different assessors at different time points.

The sample included children whose caregivers registered for the program within the first two weeks of the launch of the remedial support program in the recruited sites (N = 4,598, 49% female). As a condition for registration and continued enrollment in the remedial program, students were required to be enrolled in Lebanese public schools. The participating students' age ranged from 5 to 15 (M = 8.98, SD = 2.38); their grade level ranged from 1 to 9 in Lebanese public schools (M = 2.80, SD = 1.78); and the vast majority of them aged 12 or younger (91%) and attended 6th grade or lower (95%).

Measure

As discussed above, SRA-AR uses a short 13-item version of the assessor-report section of the Preschool Self-Regulation Assessment (Smith-Donald et al., 2007) adapted for a study in Zambia (McCoy et al., 2017). Importantly, the PSRA was originally designed to assess two aspects: (a) a performance-based assessment of self-regulation skills of pre-school children; and (b) assessors' ratings of each child on behaviors displayed during the assessment. For this study, we asked the assessors to rate the children's behavior during a one-hour-long child assessment period, which included direct assessment and self-report measures of academic, cognitive, social, and emotional skills. Each assessor was asked to report the children's behavior they observed during the entire assessment period, on items such as: "Pays attention to instructions and demonstration," "Remains in seat appropriately during the test." Each item was scored on a four-point scale, with higher scores indicating better behavioral regulation.

Analysis and Results

We present findings from the analysis conducted to provide evidence of (1) structural evidence of validity and reliability; (2) correlational evidence of validity; and (3) measurement invariance across treatment groups, gender groups, age groups, and time.

All descriptive, bivariate correlation and reliability analyses were conducted using Stata SE version 15.1, and all measurement modeling was conducted using Mplus 8.3 (Muthén & Muthén, 2014). In order to account for the structural characteristics of the data, two important specifications were made for all measurement models. First, given item response options in the measure, items were specified as categorical. Because modeling categorical responses as normally and continuously distributed can lead to an inflation of model fit statistics and biased estimation of factor loadings and standard errors, we used a weighted least squares mean and variance-adjusted (WLSMV) estimator with a probit-link function (Beauducel & Herzberg, 2006; Lei, 2009). Second, we used robust standard errors to adjust for clustering because 1) students were nested within classrooms/teachers, and classrooms/teachers within sites; and 2) it

was an effective and efficient way to model complex data when sample size at the cluster level was not small (Huang, 2016). In all models, model fits were evaluated using Hu and Bentler's (1999) criteria: RMSEA (Root Mean Square Error Of Approximation) < 0.06, CFI (Comparative Fit Index) < 0.95, TLI (Tucker–Lewis Index) < 0.95, SRMR (Standardized Root Mean Squared Residual) < 0.08. Missing data were pairwise deleted (i.e., all available information was used from all cases) to preserve the full sample (Asparouhov & Muthén, 2010). As a result, we were able to include and obtain factor scores for all children who were ever assessed for any items of CFS in the analysis regardless of missing information on specific items.

Table 1

PSRA item names and item descriptions

Item	Item description
PSRA1	Pays attention to instructions and demonstration
PSRA2	Careful, interested in accuracy
PSRA3	Sustains concentration; willing to try repetitive tasks
PSRA4	Is careless or destructive with test materials
PSRA5	Can wait during and between tasks
PSRA6	Remains in seat appropriately during the test
PSRA7	Alert and interactive; is not withdrawn
PSRA8	Cooperates; complies with requests
PSRA9	Shows pleasure in accomplishment and active task mastery
PSRA10	Confident
PSRA11	Defiant
PSRA12	Passively noncompliant
PSRA13	Modulates and regulates arousal level in self

Aim 1: Structural Evidence of Validity and Reliability

To address Aim 1, we conducted (a) exploratory and confirmatory factor analyses (EFA and CFA); and (b) estimation of internal consistency statistics (Cronbach's α and Mcdonald's ω).

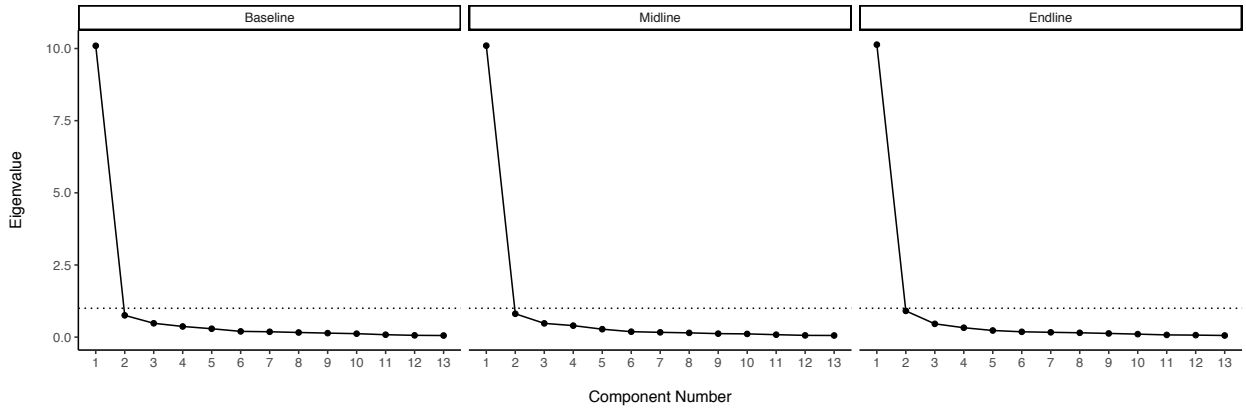
Exploratory and confirmatory factor analysis. Before beginning analyses, we randomly divided our sample in half in order to create exploratory and confirmatory samples at each time point. Exploratory samples were used to examine multiple versions of data-driven models, of which a final proposed solution was selected based on conceptual and empirical considerations. Confirmatory samples were used to test the proposed factor structure, thereby builds confidence in the stability of empirically derived exploratory factor analytic estimates (Osborn &

Fitzpatrick, 2012). CFA models with a good model fit and the same factor structure across baseline, midline, and endline were used as final models for subsequent analysis.

Then, we performed a series of EFAs to empirically explore the factor structure. The scree plots of eigenvalues suggested a 1-factor structure with elbows at the second factor and eigenvalue $<$ 1 at factor 2 (see Figure 1). Based on the one-factor structure in the original PSRA-AR measure, we also fit a 1-factor structure, and results showed that this theory-based structure of the measure also fitted well across all waves of the data. Indeed, all 13 items describe similar behaviors or general states that exemplify behavioral regulation, and they appear to measure a congruent construct of behavioral regulation.

Figure 1

Scree plot of eigenvalues from exploratory factor analysis at all waves



Using the solution from the EFA, we ran a CFA with a one-factor model. However, this model yielded unsatisfactory model fits (RMSEA = 0.087). The modification indices (MI) suggested inter-item correlations of PSRA10 ("confident") with PSRA9 ("Shows pleasure in accomplishment and active task mastery"), PSRA6 ("Remains in seat appropriately during the test ") with PSRA4 ("Is careless or destructive with test materials") and PSRA8 ("Cooperates; complies with requests") with PSRA 7 ("Alert and interactive; is not withdrawn") (see Figure 2). Given the thematic connection between these items (sense of mastery; physical control; cooperativeness) and MI statistics, we fit another CFA model adding these residual covariances, which yielded good model fits. This same final model of the baseline was tested for midline and endline and yielded a result with a good model fit. All items loaded onto the factor with high factor loadings at $\lambda > 0.70$ at baseline and midline, and $\lambda > 0.80$ at endline. See also Table 2 for descriptive statistics and Table 3 for the factor loadings of each item at all waves.

Figure 2

Factor structure displaying model parameters at all waves

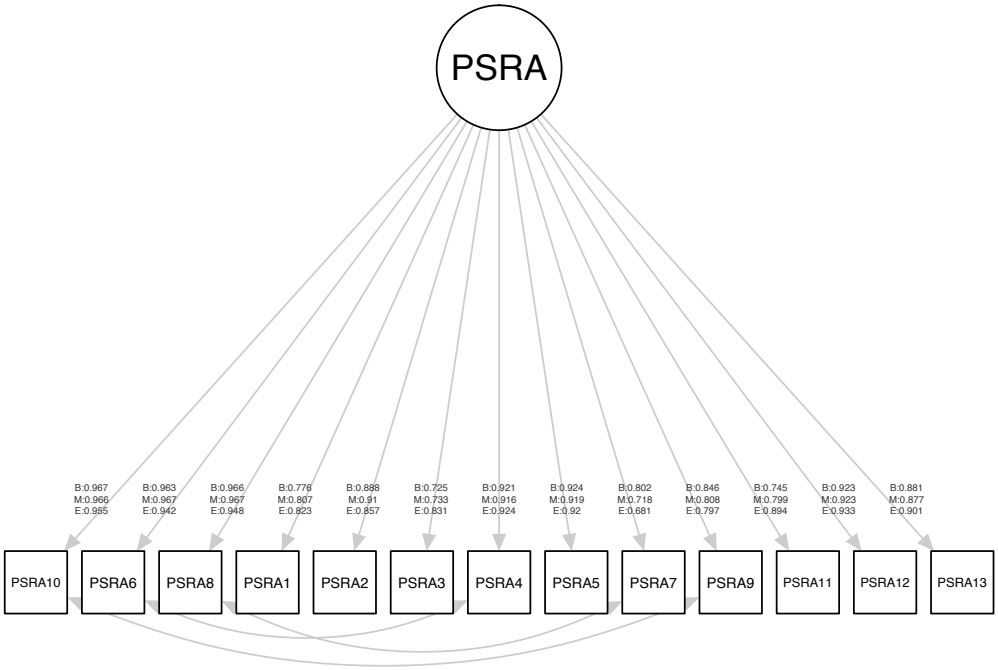


Table 2

Descriptive statistics of indicators by proposed construct

	Baseline			Midline			Endline		
	N	Mean	SD	N	Mean	SD	N	Mean	SD
PSRA1	4276	3.210	1.037	3523	3.398	0.914	3706	3.481	0.828
PSRA2	4273	3.096	1.086	3525	3.297	0.985	3706	3.393	0.895
PSRA3	4273	3.152	1.045	3523	3.343	0.924	3708	3.420	0.845
PSRA4	4267	3.664	0.745	3523	3.776	0.620	3705	3.725	0.648
PSRA5	4256	3.289	1.013	3519	3.367	0.907	3704	3.417	0.820
PSRA6	4250	3.748	0.668	3527	3.830	0.537	3704	3.784	0.568
PSRA7	4265	3.298	0.971	3521	3.451	0.839	3694	3.496	0.790
PSRA8	4268	3.361	0.959	3521	3.508	0.837	3698	3.546	0.777
PSRA9	4271	3.019	0.877	3524	3.017	0.848	3704	2.996	0.842
PSRA10	4252	2.945	0.995	3524	3.022	0.921	3705	3.088	0.862
PSRA11	4232	3.492	0.887	3521	3.618	0.785	3687	3.657	0.727
PSRA12	4272	3.435	0.903	3525	3.549	0.812	3703	3.591	0.737
PSRA13	4265	3.562	0.826	3520	3.684	0.675	3704	3.726	0.636

Table 3

Factor loadings of the SRA-AR at all waves from the CFA final model[2]

		Baseline			Midline			Endline		
		b	*SE*	*p*	*b*	*SE*	*p*	*b*	*SE*	*p*
Preschool children's self-regulation										
(Baseline α = .955, ω = .978; Midline α = .952, ω = .977; Endline α = .953, ω = .977)										
PSRA1	Pays attention to instructions and demonstration	0.967	0.003	0	0.966	0.003	0	0.955	0.004	0
PSRA2	Careful, interested in accuracy	0.963	0.003	0	0.967	0.003	0	0.942	0.005	0
PSRA3	Sustains concentration; willing to try repetitive tasks	0.966	0.003	0	0.967	0.003	0	0.948	0.005	0
PSRA4	Is careless or destructive with test materials	0.776	0.016	0	0.807	0.017	0	0.823	0.014	0
PSRA5	Can wait during and between tasks	0.888	0.01	0	0.91	0.008	0	0.857	0.01	0
PSRA6	Remains in seat appropriately during the test	0.725	0.022	0	0.733	0.021	0	0.831	0.015	0
PSRA7	Alert and interactive; is not withdrawn	0.921	0.006	0	0.916	0.007	0	0.924	0.007	0
PSRA8	Cooperates; complies with requests	0.924	0.006	0	0.919	0.008	0	0.92	0.007	0
PSRA9	Shows pleasure in accomplishment and active task mastery	0.802	0.013	0	0.718	0.019	0	0.681	0.017	0
PSRA10	Confident	0.846	0.009	0	0.808	0.013	0	0.797	0.011	0
PSRA11	Defiant	0.745	0.019	0	0.799	0.018	0	0.894	0.009	0
PSRA12	Passively noncompliant	0.923	0.006	0	0.923	0.008	0	0.933	0.007	0
PSRA13	Modulates and regulates arousal level in self	0.881	0.01	0	0.877	0.012	0	0.901	0.01	0

[2] The table displays standardized coefficients.

Internal consistency of subscales. For assessing internal consistency, Cronbach's alpha (α) of each latent factor within each data collection time point was calculated. We also assessed McDonald's omega (ω; Hayes & Coutts, 2020; McDonald, 1999) of each latent factor as a more general reliability estimate that does not assume equal factor loadings (i.e., tau-equivalence). The recommendation from the contemporary literature of assessing reliability for unidimensional measures assuming unequal factor loadings, like SRA-AR, is to avoid α and use ω (Revelle and Zinbarg 2009; Zinbarg et al. 2005). Therefore, we report both reliability statistics but focus on the interpretation of ω. While there are no definitive and universal guidelines for interpreting α and ω, $\alpha > 0.7$ is generally accepted as acceptable/high reliability, and Nájera Catalán (2019) suggests a higher standard for $\omega > 0.8$ as excellent evidence of internal consistency.

Table 3 presents both the unweighted (Cronbach's α) and the weighted (McDonald's ω) internal consistency estimates of the SRA-AR scale. Overall, the SRA-AR has high internal consistency at all waves (Baseline α = .955, ω = .978; Midline α = .952, ω = .977; Endline α = .953, ω = .977).

Aim 2: Correlational Evidence of Validity

To address Aim 2, we examined the correlation among SRA-AR scores across time using factor scores. In addition, we present evidence of correaltions between SRA-AR and other related constructs reported in Kim et al. (2020) which used the data of the subsample reported here.

Correlations across time. As seen in Table 4, the correlations between SRA-AR across baseline, midline, and endline were moderate, r = .396 - .467. This indicates that there was a degree of stability in the measurement of children's behavioral regulation skills such that between 10%-15% of the variability in behavioral regulation skills at midline and endline are explained by behavioral regulation skills measured at a previous time point (i.e., baseline). This also suggests that the Syrian children's behavioral regulation skills captured by SRA-AR changed to a certain extent but remained generally stable over the course of the school year.

Table 4

Construct-level correlations across waves

	Baseline	Midline	Endline
1. PSRA_1	--	--	--
2. PSRA_2	0.467***	--	--
3. PSRA_3	0.396***	0.436***	--

Correlations between SRA-AR and other related constructs. Given the importance of behavioral regulation skills to successfully adjust to and learn in schools (Duncan et al, 2017; Ursache et al, 2012), it is expected to be positively associated with academic outcomes and other skills that are related to academic skill. Authors of this report used a subsample of the larger cRCT study (the data on SRA-AR reported here are from the full sample of the study) and examined how various social and emotional skills—including behavioral regulation measured by SRA-AR—are associated with academic outcomes. As reported in Kim et al. (2020), and as expected, behavioral regulation measured using SRA-AR was positively correlated with Syrian refugee children's key cognitive skills necessary for learning – working memory (r = .27) and inhibitory control ($r = .10$) – as well as their academic skills, literacy ($r = .58$) and numeracy ($r = .58$). It was not related to the internalizing symptoms (Table 5). See details on the sample and measures of other related constructs in Kim et al., (2020).

Table 5

Correlations between SRA-AR and other related constructs

	1	2	3	4	5	6
1 Working memory	-					
2 Inhibitory control	0.21***	-				
3 Internalizing symptoms	-0.12*	-0.16**	-			
4 Behavioral regulation (SRA-AR)	0.27***	0.10**	-0.11	-		
5 Literacy	0.49***	0.31***	-0.20***	0.58***	-	
6 Numeracy	0.41***	0.27***	-0.23***	0.62***	0.93***	-

Note. This table originally appeared in: Kim, H. Y., Brown, L., Tubbs Dolan, C., Sheridan, M., & Aber, J. L. (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

Aim 3: Evidence of Measurement Invariance

To address Aim 3, we conducted (1) measurement invariance tests across treatment, gender, and age groups in each wave; and (2) longitudinal invariance testing across baseline, midline, and endline. 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 times (Reise, Widaman, & Pugh, 1993). If a measure operates or is understood differently in different groups, then one should not compare group differences on observed scores (Glanville & Wildhagen, 2007). For example, without evidence of measurement invariance, one should not compare boys' and girls'

self-regulation; compare this construct with and without access to SEL interventions; or track changes in students' self-regulation over time.

For each set of analyses, we tested for levels of measurement invariance by fitting a series of nested models in which we progressively constrained the model parameters to equality across groups/time points. Specifically, we fit models within each time point and then across time points to test the equality of 1) the factor structure in treatment and control groups and time points (configural invariance); 2) the factor loadings across groups/time points (metric invariance); and 3) the item intercepts or thresholds across groups/time points (scalar invariance) (Millsap, 2012; Gregorich, 2006). We assessed the relative fit of each of these models 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). If the imposition of equality constraints did not provide a significant decrement of model fit, we concluded that the hypothesis of invariance was supported.

Treatment invariance. We found evidence of scalar invariance at all waves between treatment and control groups (see Table 6 for model fits). This means that the latent factors across two different treatment groups measure equivalent constructs, and therefore we can directly compare treatment and control group students on the same SRA-AR scale, without bias.

Gender and age measurement invariance. We found that SRA-AR was scalar invariant at all waves across gender and age groups (see Tables 7 & 8 for model fits), suggesting that we can compare mean differences by gender and age on the SRA-AR constructs without bias due to child gender or age.

Invariance across time. As shown in Table 8, a series of longitudinal invariance models were tested to confirm that changes from baseline to midline, and midline to endline, of the same construct can be estimated. Model fit difference between configural, metric, and scalar models suggested the factor structure, loadings, and thresholds of the items were invariant from baseline to endline. In other words, we found no significant difference in the item and measure functioning across waves, and we can compare baseline, midline, and endline scores on these constructs as assessed using the SRA-AR.

Table 6

Treatment group invariance model fit

k	χ2	*df*	*p*	Δχ2	*df*	*p*	CFI	TLI	RMSEA	SRMR
Baseline										
165	1166.042	186	0	NA	NA	NA	0.993	0.991	0.061	0.032
141	721.321	210	0	35.441	24	0.0621	0.996	0.996	0.041	0.033
65	805.419	286	0	116.252	76	0.002	0.996	0.997	0.036	0.034
Midline										
165	794.104	186	0	NA	NA	NA	0.994	0.992	0.053	0.032
141	610.231	210	0	76.506	24	0	0.996	0.996	0.04	0.038
65	708.857	286	0	144.255	76	0	0.996	0.997	0.035	0.039
Endline										
165	1153.729	186	0	NA	NA	NA	0.993	0.991	0.065	0.036
141	829.931	210	0	70.955	24	0	0.996	0.995	0.049	0.039
65	862.256	286	0	93.112	76	0.0887	0.996	0.997	0.04	0.039

Table 7

Gender group invariance model fit

k	χ2	*df*	*p*	Δχ2	*df*	*p*	CFI	TLI	RMSEA	WRMR
Baseline										
110	1418.36	124	0	NA	NA	NA	0.992	0.99	0.07	2.311
98	793.164	136	0	22.895	12	0.0286	0.996	0.995	0.048	2.405
60	824.316	174	0	55.903	38	0.0306	0.996	0.996	0.042	2.448
Midline										
110	952.611	124	0	NA	NA	NA	0.994	0.993	0.062	2.011
98	548.914	136	0	15.23	12	0.2291	0.997	0.997	0.041	2.088
60	573.888	174	0	36.755	38	0.527	0.997	0.998	0.036	2.117
Endline										
110	1511.341	124	0	NA	NA	NA	0.992	0.989	0.078	2.764
98	916.541	136	0	12.879	12	0.3779	0.995	0.995	0.056	2.808
60	886.69	174	0	31.882	38	0.7472	0.996	0.996	0.047	2.827

Table 8

Age group invariance model fit

k	χ2	*df*	*p*	Δχ2	*df*	*p*	CFI	TLI	RMSEA	WRMR
Baseline										
220	1556.827	248	0	NA	NA	NA	0.992	0.99	0.07	2.639
184	984.847	284	0	91.29	36	0	0.996	0.995	0.048	3.001
70	1132.413	398	0	183.19	114	0	0.996	0.996	0.042	3.145
Midline										
220	1122.394	248	0	NA	NA	NA	0.993	0.992	0.063	2.301
184	830.485	284	0	100.969	36	0	0.996	0.996	0.047	2.765
70	971.078	398	0	186.958	114	0	0.996	0.997	0.04	2.942
Endline										
220	1696.88	248	0	NA	NA	NA	0.991	0.988	0.079	3.001
184	1179.158	284	0	95.875	36	0	0.994	0.994	0.058	3.318
70	1249.416	398	0	160.722	114	0.0026	0.995	0.996	0.048	3.454

Table 9

Longitudinal invariance model fit

k	χ2	*df*	*p*	Δχ2	*df*	*p*	CFI	TLI	RMSEA	SRMR
168	1545.067	690	0	NA	NA	NA	0.996	0.996	0.016	0.031
144	2026.944	714	0	383.161	24	0	0.995	0.994	0.02	0.039
68	2194.967	790	0	366.441	76	0	0.994	0.994	0.02	0.04

Conclusion

SRA-AR was developed based on the original PSRA-AR measure to assess enumerators' perceptions of Syrian refugee students' behavioral regulation skills during an assessment. It was used to evaluate the impact of access to non-formal, SEL-infused remedial support programming among Syrian refugee children in Lebanon. Evidence indicates that SRA-AR holds promise for use as a program evaluation measure, with evidence of validity based on its internal structure and consistency, correlational patterns, and measurement invariance across treatment groups, gender, age groups, and time.

First, measures used for program evaluation purposes must have strong structural evidence of validity: evidence that scores on the measure can be interpreted as capturing key dimensions of behavioral regulation. Factor analyses of SRA-AR suggested a one-factor solution, and all items had high loadings on to this factor at all waves. This provides strong evidence for the structural evidence of validity of the measure.

In terms of internal consistency, data from program evaluation measures must be highly reliable, as measurement error can attenuate the ability to detect program impact (Raudenbush & Sadoff, 2008). All of the empirically derived subscales had high internal consistency, indicating that enumerators generally gave consistent ratings on items on SRA-AR (alpha > 0.95).

Second, correlational patterns of the SRA-AR across time and with other related constructs provides evidence that the SRA-AR captures relevant information for behavioral regulation that are relatively stable over time and associated with theoretically related constructs in expected directions. Specifically, correlation across time were moderate between waves, indicating children's behavioral regulation may fluctuate moderately during the school year. While these correlations are not very high, they are aligned with studies conducted in the U.S. suggesting that SEL constructs tend to be more strongly influenced by contextual factors and likely to be time-varying, compared to academic skills that tend to be highly stable over time (Soland et al. 2019). These moderate levels of correlations are also notable considering the reporter effect (i.e., the different reporters/assessors) across time points, which is likely to lead to lower correlation. In addition, the strong correlation between behavioral regulation and academic skills – as well as small but positive correlations with relevant cognitive skills such as working memory and inhibitory control – strengthens the evidence of validity.

Third, data from program evaluation measures should also provide evidence that the measure functions are well understood in the same way by children in different treatment and demographic groups, as well as over time. This criterion is known as measurement invariance. Establishing the measurement invariance of an assessment used in a rigorous program impact evaluation enables us to confidently assess whether children's skills are improving or declining over time – and whether such changes are the result of our SEL programming (Halpin et al., 2019; Halpin & Torrente, 2014). SRA-AR has strong evidence on:

- Longitudinal invariance, suggesting that the measure can be used to directly assess growth over time on scores from the SRA-AR.
- Gender invariance, suggesting that scores from the derived SRA-AR scale can be used to capture meaningful gender differences in behavioral regulation. This piece of evidence indicates that the measure is not biased when deriving gender mean differences.
- Age invariance, suggesting that scores from the derived SRA-AR scale can be used to capture meaningful age group differences in behavioral regulation. This piece of evidence indicates that the measure is not biased when deriving mean differences by age.
- Treatment invariance, suggesting that scores from the derived of the SRA-AR scale can be used to capture meaningful treatment group differences in behavioral regulation. This piece of evidence indicates that the measure is not biased when deriving mean differences by treatment group.

Limitations of the Measure

The SRA-AR has strong evidence that the resulting data can be used to make valid inferences about Syrian refugee children's behavioral regulation skills. The measure is a performance-based observation measure, which can be used to evaluate discrete and specific components of children's performance on specific tasks, including how the task was approached. In the present case, children's behavioral regulation was reported by the assessor (i.e., data collector/enumerator) based on children's behavioral performance during a one-hour data collection session. Thus scores on this measure represent children's behavioral regulation skills during one assessment session with one assessor. Given this, one of the largest limitations of the measure is that it does not represent any given child's global behavioral regulation competency across multiple types of situations and contexts (i.e., during classroom activities, in collaboration with peers, at home). In addition, due to limited resources and training time, we did not collect information on enumerators' inter-rater reliability, or the extent to which enumerators provide consistent ratings of children's behavioral regulation skills. While relatively

stable correlation across time—assessed by different enumerators—indicates some level of inter-rater reliability, we do not have data on concurrent observation of different enumerators of the same child. For the current study, SRA-AR was included as "easy add-on" to existing assessment protocol and focused on its utility, efficiency, and feasibility of field use. Therefore, the training of the SRA-AR focused on ensuring common understanding of the meanings of each items and response options; and did not include rigorous observation training protocol typically required to achieve interrater reliability of an observation measure (e.g., consensus coding with experts, IRR test as a part of training). Future studies using this measure should consider developing a more comprehensive training protocol and processes of establishing and measuring interrater reliability as a part of training and/or data collection, in order to strengthen the reliability of the measure.

Recommendations for the Use of the SRA-AR

While the evidence provided in this study largely supports the use of SRA-AR for evaluation purposes with Syrian refugee children in Lebanon, a few implications should be noted when researchers and practitioners consider the use of SRA-AR for their own purposes. Any extensions of the use of SRA-AR are not recommended without adaptation and a re-evaluation of the psychometric properties of the measure. With adequate empirical evidence and once SRA-AR is deemed appropriate for the setting and purpose, we recommend a set of strategies and future directions to ensure that children's behavioral regulation is accurately interpreted:

1. Adaptation and translation of the SRA-AR can benefit from cognitive interviews to ensure cultural and linguistic fit of the wording of the items. In addition, given that the SRA-AR items provide discrete response options to anchor response patterns, it may be informative to explicitly evaluate how differently each response options are perceived by the respondents (assessors) and how frequently they think they would observe each response options. Adaptations of the items and response options will not only ensure validity of the measure but also improve the distribution of the item responses to prevent ceiling or floor effects.
2. If adequate time and resources are available, we recommend administering the SRA-AR at multiple occasions (e.g., at a different time of the day, across multiple days) and in different school contexts that require goal-oriented behaviors (e.g., during classroom instruction; individual work; assessment). Such variation would be desirable to capture a more global understanding of children's underlying behavioral regulation skills and to minimize measurement errors due to the contextual factors, e.g., specific tasks, mood, or time of the day.

3. In order to promote the consistency of behavioral regulation measured across time and across children, it is important to standardize the environmental factors that may come into effect during the data collection session. For example, if data are collected in a distracting environment for one child and not for another, behavioral regulation is likely to vary based on what other distractions are present in the environment of the performance-based period.
4. Explicit assessor training to fill out the survey can ensure the validity and reliability of assessor reports. Assessors typically do not have the experience or training to carefully observe and report discrete behaviors of the children, and sometimes may have varying understanding of certain concepts describing children's behavioral regulation. Therefore, establishing common understandings across the assessors of the meaning of not just the items but also the response options presented in SRA-AR for the concepts each of the items are intended to capture will be necessary to ensure its reliability and validity.

References

Asparouhov, T., & Muthén, B. (2010). Weighted least squares estimation with missing data. Mplus Technical Appendix.

Beauducel, A., & Herzberg, P. Y. (2006). On the performance of maximum likelihood versus means and variance adjusted weighted least squares estimation in CFA. Structural Equation Modeling: A Multidisciplinary Journal, 13(2), 186–203. https://doi.org/10.1207/s15328007sem1302_2

Blair, C. (2010). Stress and the development of self-regulation in context. Child Development Perspectives, 4(3), 181–188. https://doi.org/10.1111/j.1750-8606.2010.00145.x

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

Connor-Smith, J. K., Compas, B. E., Wadsworth, M. E., Thomsen, A. H., & Saltzman, H. (2000). Responses to stress in adolescence: Measurement of coping and involuntary stress responses. Journal of Consulting and Clinical Psychology, 68(6), 976.

Dodge, K. A., Pettit, G. S., Bates, J. E., & Valente, E. (1995). Social information-processing patterns partially mediate the effect of early physical abuse on later conduct problems. Journal of Abnormal Psychology, 104(4), 632–643. https://doi.org/10.1037/0021-843X.104.4.632

Duncan, R. J., McClelland, M. M., & Acock, A. C. (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.

Glanville, J. L., & Wildhagen, T. (2007). The measurement of school engagement: Assessing dimensionality and measurement invariance across race and ethnicity. Educational and Psychological Measurement, 67(6), 1019–1041. https://doi.org/10.1177/0013164406299126

Gregorich, S. E. (2006). Do self-report instruments allow meaningful comparisons across diverse population groups? Testing measurement invariance using the confirmatory factor analysis framework. Medical Care, 44(11 Suppl 3), S78–S94. https://doi.org/10.1097/01.mlr.0000245454.12228.8f

Halpin, P. F., & Torrente, C. (2014). Measuring critical education processes and outcomes: illustration from a cluster randomized trial in the Democratic Republic of the Congo. In Society for Research on Educational Effectiveness. Society for Research on Educational Effectiveness. http://eric.ed.gov/?id=ED562783

Halpin, P. F., Wolf, S., Yoshikawa, H., Rojas, N., Kabay, S., Pisani, L., & Dowd, A. J. (2019). Measuring early learning and development across cultures: Invariance of the IDELA across

five countries. Developmental Psychology, 55(1), 23–37. https://doi.org/10.1037/dev0000626

Hayes, A. F., & Coutts, J. J. (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

Henrich, J., Heine, S. J., & Norenzayan, A. (2010). Most people are not WEIRD. Nature, 466(7302), 29–29. https://doi.org/10.1038/466029a

Hu, L., & Bentler, P. M. (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

Huang, F. L. (2016). Alternatives to multilevel modeling for the analysis of clustered data. The Journal of Experimental Education, 84(1), 175–196. https://doi.org/10.1080/00220973.2014.952397

Khamis, V. (2000). Political violence and the Palestinian family: Implications for mental health and well-being (pp. xv, 144). Haworth Maltreatment and Trauma Press/The Haworth Press.

Linares, L. O., Heeren, T., Bronfman, E., Zuckerman, B., Augustyn, M., & Tronick, E. (2001). A mediational model for the impact of exposure to community violence on early child behavior problems. Child Development, 72(2), 639–652. https://doi.org/10.1111/1467-8624.00302

Lei, P.-W. (2007). 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

McCoy, D. C., Zuilkowski, S. S., Yoshikawa, H., & Fink, G. (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, R. P. (1999). Test theory: A unified treatment (pp. xi, 485). Lawrence Erlbaum Associates Publishers.

Millsap, R. E. (2011). Statistical approaches to measurement invariance (pp. xii, 355). Routledge/Taylor & Francis Group.

Muthén, L. K., & Muthén, B. (2018). Mplus. The comprehensive modelling program for applied researchers: User's guide, 5.

Nájera Catalán, H. E. (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

Osborne, J. W., & Fitzpatrick, D. C. (2012). Replication analysis in exploratory factor analysis: What it is and why it makes your analysis better. Practical Assessment, Research & Evaluation, 17(15). https://doi.org/10.7275/h0bd-4d11

Pandey, A., Hale, D., Das, S., Goddings, A.-L., Blakemore, S.-J., & Viner, R. (2017). Effectiveness of universal self-regulation-based interventions to improve self-regulation, and effects on distant health and social outcomes in children and adolescents: A systematic review and meta-analysis. The Lancet, 390, S66. https://doi.org/10.1016/S0140-6736(17)33001-5

Raudenbush, S. W., & Sadoff, S. (2008). Statistical inference when classroom quality is measured with error. Journal of Research on Educational Effectiveness, 1(2), 138–154. https://doi.org/10.1080/19345740801982104

Reise, S. P., Widaman, K. F., & Pugh, R. H. (1993). Confirmatory factor analysis and item response theory: Two approaches for exploring measurement invariance. Psychological Bulletin, 114(3), 552–566. https://doi.org/10.1037/0033-2909.114.3.552

Revelle, W., & Zinbarg, R. E. (2008). Coefficients Alpha, Beta, Omega, and the glb: Comments on Sijtsma. Psychometrika, 74(1), 145. https://doi.org/10.1007/s11336-008-9102-z

Robson, D. A., Allen, M. S., & Howard, S. J. (2020). Self-regulation in childhood as a predictor of future outcomes: A meta-analytic review. Psychological Bulletin, 146(4), 324–354. https://doi.org/10.1037/bul0000227

Smith-Donald, R., Raver, C. C., Hayes, T., & Richardson, B. (2007). Preliminary construct and concurrent validity of the Preschool Self-Regulation Assessment (PSRA) for field-based research. Early Childhood Research Quarterly, 22(2), 173–187. https://doi.org/10.1016/j.ecresq.2007.01.002

Soland, J., Kuhfeld, M., Wolk, E., & Bi, S. (2019). Examining the state-trait composition of social-emotional learning constructs: Implications for practice, policy, and evaluation. Journal of Research on Educational Effectiveness, 12(3), 550–577. https://doi.org/10.1080/19345747.2019.1615158

Ursache, A., Blair, C., & Raver, C. C. (2012). The promotion of self-regulation as a means of enhancing school readiness and early achievement in children at risk for school failure. Child Development Perspectives, 6(2), 122–128. https://doi.org/10.1111/j.1750-8606.2011.00209.x

Zinbarg, R. E., Revelle, W., Yovel, I., & Li, W. (2005). Cronbach's, α Revelle's β and McDonald's ω H: Their relations with each other and two alternative conceptualizations of reliability. Psychometrika, 70(1), 123–133. https://doi.org/10.1007/s11336-003-0974-7