3EA TIES-IRC Measurement Tools: Children's Holistic Learning and Development (CHILD) and Program Implementation Quality (PIQ)

Technical Memo

Introduction

As part of the Education in Emergencies: Evidence for Action (3EA) initiative, New York University's Global TIES for Children (TIES/NYU) and the International Rescue Committee (IRC) are leading a group of research-practice-policy partnerships to develop, adapt, and test a set of measurement tools to assess critical dimensions of program implementation quality (PIQ) and children's learning and holistic development (CHILD) in crisis contexts. In this technical memo – oriented to researchers and technical staff at non-governmental, multi-lateral, and intergovernmental organizations – we provide information about the assessment tools our own TIES-IRC partnership is working to develop, adapt and test. Specifically, we:

- Describe the purposes for which we are developing and adapting assessments, as well as the types of assessments we are testing, where, and with whom
- Explain the types of evidence we are currently working to generate to understand the extent to which our assessments are providing accurate, meaningful, and comparable data
- Provide for interested parties best practices on how to cite, share, and communicate interest in using this set of tools

Background

Developing tools that meaningfully and accurately assess children's skills and competencies—and the key aspects of the environment that shape them—is demanding under the best of circumstances, given that children's development is a complex process that occurs dynamically and transactionally over time. In crisis contexts, we face additional challenges given the current lack of understanding of how children develop and learn within and across cultures and contexts: 95% of what we know about children's development is based on research with just 5% of the world's children, those that live in White, Educated, Industrialized, Rich, and Democratic (WEIRD) contexts.[1] And while caregivers, teachers, and policymakers across contexts may broadly agree on the skills and competencies critical for children's long-term success[2], how such skills are named, defined, manifested, operationalized, and/or prioritized differs according to the context. Designing and adapting assessments that we know capture meaningful and accurate information

[1] Henrich, Heine, and Norenzayan, "The Weirdest People in the World?"
[2] Learning Metrics Task Force, "Toward Universal Learning: Recommendations from the Learning Metrics Task Force"; Torrente, Alimchandani, and Aber, "International Perspectives on Social and Emotional Learning."

about children's holistic learning and development and program implementation quality in crisis contexts takes time and resources – something that is in short supply in the contexts in which we work.

Given these challenges, we encourage a culture of sharing measurement tools and data sets across stakeholder groups, and of transparency around how the tools are working (or not). This norm will increase the chances that the limited available resources are well-used and that the urgent need for high-quality data in crisis contexts is efficiently met. At the same time, it is important to note that a tool is just a tool: the extent to which it provides high-quality data is largely determined by the decisions made about what it is used for, how it is used, where it is used, and what it is intended to assess. To facilitate technical stakeholders' understanding of the purposes for which our TIES-IRC partnership tools are used for, as well as what they are intended to assess and where, we turn now to an overview of the "what, where, and how" of 3EA TIES-IRC measurement tools.[3]

Part 1: The What, Where, and How of 3EA TIES-IRC Measurement Tools

What are we using measurement tools for?
Different measurement tools generate data that can be used for many purposes. These include:

1) **Describing and comparing** children's learning and development at the population level to identify areas of special need and ensure accountability
2) **Tracking** individual children's formative skill development to provide individualized scaffolding and support
3) **Screening** children for developmental delays and mental health difficulties
4) **Monitoring** the implementation and quality of programming and **evaluating** the impact of programming
5) **Contributing** to hypothesis testing about how neurobiological, cognitive, social-emotional, and ecological factors interact to shape children's development.

The purposes for which the data will be used should critically guide the design/adaptation of the measure, its implementation, and how the resulting data analyzed.[4]

For example, many governments and global institutions are interested in population-level monitoring that describes at a school, community, sub-regional, and/or national aggregate level how children are learning and developing over time (Purpose 1, above). Given that such efforts require cost- and time-intensive representative or Census-based samples, tools used for population-level monitoring are often short and designed to provide a snapshot of the high-level skills and competencies

[3] We use the terms "measure," "assessment," and "measurement tool" interchangeably throughout the document to refer to instruments that can be used to collect data about children's holistic learning and development and program implementation and quality. These instruments can take many forms – from Likert-scale surveys to participatory open-ended responses to observation rubrics to performance-based games – that shape the purposes for which the resulting data can be used.
[4] Fernald et al., "Toolkit for Measuring Early Childhood Development in Low- and Middle-Income Countries."

relevant for policymakers. The resulting data are then commonly analyzed to identify and compare across groups the percentage of children, classrooms, or schools that are meeting goals—what is known as an indicator—and to make decisions about how to allocate resources to better support the attainment of such goals

TIES/NYU and the IRC are currently developing, adapting, and testing measurement tools first and foremost to assess the impact of education programming on children's learning and development through experimental or quasi-experimental research designs, and to monitor and improve the quality of program implementation (Purpose 4, above). We will also use this information to contribute to scientific knowledge about how children develop within different cultural contexts (Purpose 5, above). These purposes then have several implications for how our research-practice partnership designs, adapts, and implements measurement tools, and analyzes the resulting data:

- Design. To capture the extent to which programs are impacting children's holistic learning and development, both our CHILD and PIQ measures tend to be more fine-grained and detailed than most population-level measures.
- Implementation. Given such detail—which can increase the complexity and cost of assessment—our CHILD measures are currently designed to be collected by trained external enumerators. Our PIQ tools, however, can be collected by program staff and teachers in crisis contexts as part of routine monitoring activities.
- Analysis. Our analysis to date of CHILD and PIQ measures has focused on examining the reliability, the validity, the sensitivity to change, and the comparability of scores resulting from the measurement tools.[5] Given the lack of availability of a population-sampling frame, we have not established benchmarks and corresponding indicators for these measurement tools.

What are we measuring?

The IRC designs and implements education programming for children in crisis contexts that focuses on social-emotional learning (SEL; see text box below). Our CHILD measurement tools were selected in accordance with the programs' theories of change to assess the range of social-emotional, cognitive, and academic competencies that the programs are hypothesized to improve.[6] Our PIQ tools are designed to assess exposure (how much of the programs was delivered), fidelity (the degree to which the program was delivered as intended) and participation (how engaged participants were).

[5] Tubbs Dolan, "The Strengths and Difficulties of the Strengths and Difficulties Questionnaire: Cross-National Measurement of Children's Social-Emotional Well-Being in Crisis-Affected Contexts"; Tubbs Dolan, "Improving the Quality of Education in the Syrian Refugee Response Region: Insights from Implementation Science"; Godfrey et al., "Cross-National Measurement of School Learning Environments"; Seidman et al., "Assessment of Pedagogical Practices and Processes in Low- and Middle Income Countries: Findings from Secondary School Classrooms in Uganda"; Wolf, Halpin, and Yoshikawa, "Evaluating the Factor Structure and Invariance of the International Development and Learning Assessment (IDELA) across Five Countries."

[6] Aber et al., "Promoting Children's Learning and Development in Conflict-Affected Countries"; Aber et al., "Impacts After One Year of 'Healing Classroom' on Children's Reading and Math Skills in DRC"; Torrente et al., "Improving the Quality of School Interactions and Student Wellbeing"; Education in Emergencies: Evidence for Action, "IRC Healing Classrooms Retention Support Programming Improves Syrian Refugee Learning in Lebanon."

What are social-emotional competencies?

Social-emotional competencies are the skills, attitudes, knowledge, and behaviors that help children effectively and positively manage daily responsibilities and challenges. Also referred to as non-cognitive skills, life skills or 21st century skills, social-emotional competencies fall along three inter-related and dynamic domains of child development:[7]

- **Emotional processes:** for example, emotion recognition and regulation, empathy, and perspective taking
- **Social and interpersonal processes:** for example, interpreting others' behavior, communicating clearly, respecting others
- **Cognitive processes:** for example, identification of the connection between actions and consequences, identifications of alternative ways to solve conflicts, working memory, inhibiting inappropriate responses, attention control, and higher-level executive functioning skills

Caregivers, teachers, communities, and policies in cultures around the world focus on ensuring what we here term children's "social-emotional competencies", but what they may describe as "citizenship competencies" or "moral education."[8] As noted above, however, local norms define the extent to which specific behaviors, values, and attitudes are desirable and appropriate, as well as expectations about how and where social-emotional competencies are learned. Information about children's social-emotional skills and the settings that shape them, then, should be collected using assessments that reflect local priorities and that capture local understandings and manifestations of social-emotional skills and practices.

How are we ensuring measurement tools accurately measure children's holistic learning and development and program implementation quality within cultural contexts?

When collecting CHILD and PIQ data in a context in which there are limited validated measurement tools available, stakeholders have a choice between adapting an existing measure or creating a new measure. If the former, there exists a spectrum of adaptation procedures to reduce systematic bias due to cultural practices and understandings: (1) adoption, in which a test is directly translated and used in a new context; (2) adaptation, in which test items and procedures are adapted for local understanding and norms; and (3) assembly, in which the degree of modification and/or pooling items and procedures from multiple sources results in a new measure. Creating a new measure, in which the content and method of administration are specifically designed and tailored to the local context, often results in the most contextually valid measure; however, it is more time- and resource- intensive and can limit comparability of data across contexts.

[7] Jones and Bouffard, "Social and Emotional Learning in Schools."
[8] Torrente, Alimchandani, and Aber, "International Perspectives on Social and Emotional Learning," 2015.

In 3EA to date, our CHILD assessments are adapted or assembled from existing measurement tools (see attached for a table of the assessments and their original authors). In doing so, we have followed recommended procedures for the adaptation and assembly process, including:[9]

1. Translation of assessment tools and instructions, and corrections of the translated version
2. Reviewing the measure content and administration procedures for appropriateness with local field staff and community members
3. Cognitive pre-tests, pilot testing and analysis
4. Iterative revisions based on pilot and baseline results

Our PIQ tools to date are both adaptations of existing measures as well as new measures created by local and global IRC staff. In collaboration with TIES/NYU, these tools are currently undergoing a similar process of iterative adaptation and testing.

Where are we testing CHILD and PIQ measurement tools?
As described in the attached tables, TIES/NYU and the IRC are currently collaborating on testing CHILD and PIQ measurement tools:

- In Lebanon, with ~3,700 Syrian refugee children ages 5 – 15 enrolled in IRC non-formal tutoring programs in Bekaa and Akkar; and
- In Niger, with ~2,000 Nigerian refugee children, Nigerien internally displaced children, and Nigerien local children ages 5 – 15 enrolled in IRC non-formal tutoring programs in the Diffa region

Our collaborative measure development work in these two countries is conducted in the context of randomized control trials (RCTs) evaluating the impact of adding targeted SEL strategies (school year 2016-2017 in Lebanon and Niger; school year 2017-2018 in Niger) and explicit SEL instruction (school year 2017-2018 in Lebanon) to non-formal after-school programs intended to support children's retention in formal school systems. We are also conducting a mixed methods study of the PIQ tools with teachers in the Bo region of Sierra Leone.

The IRC is additionally working:

- In Nigeria, to test this suite of CHILD and PIQ tools in the context of two RCTs in the regions of Yobe and Borno. The first RCT is conducted with 80 teachers and 2,880 children, ages 9-14, to evaluate the effect of different models of professional development for teachers implementing an accelerated learning program in non-formal learning centers. The second RCT is being conducted to evaluate the effect of a tutoring intervention on students' academic and SEL outcomes with 2,373 children in formal schools.
- In the Kurdish Republic of Iraq, to develop and test new PIQ tools to assess the implementation of different components of their teacher professional development program.

[9] Fernald et al., "Toolkit for Measuring Early Childhood Development in Low- and Middle-Income Countries."

Part 2: Sharing principles

We are committed to ensuring our adaptations of measurement tools are open source. We are also committed to providing potential users with the information – the evidence of reliability, validity, comparability, and sensitivity to change (see Part 3, below) – that is necessary for the appropriate selection and use of measurement tools. Until that analysis is complete in mid-2019, we ask parties interested in using our versions of the measurement tools to adhere to the following principles:

- Alert us of your interest in using the tools. We are happy to share versions of the tools on a one-on-one basis and to orient potential users to the measures and what we have learned to date about their psychometric properties.
- Share any further adaptations of our tools and information about where and how the data was collected. This will help us to track the conditions under which the measure has been used, ultimately creating an evidence base for tool use.
- Let us know if others are interested in these tools. We ask at this point that you do not forward the adapted tools beyond your own team; rather, that you put us in touch directly with any interested parties.
- Cite the tools according to their original and adapted versions. We provide tables below with these suggested citations.

Part 3: What criteria are we using to evaluate our measurement tools?

Under the best of circumstances, establishing the psychometric properties of a measurement tool—its reliability, validity, comparability and sensitivity to change—while also ensuring its feasibility of use requires negotiation between scientific rigor and the realities of the context in which the measure is being used. These terms are defined below drawing primarily on conceptualizations of these criteria provided in the Standards for Educational and Psychological Assessment.[10] We also include a discussion of common challenges to meeting these criteria – and the implications for researchers, practitioners, and policy makers for not meeting these criteria - in crisis contexts.

Reliability

Most generally, reliability refers to the extent to which scores on an assessment tool are consistent, coherent, and precise: Do the different items within each assessment provide a consistent picture of a students' ability or do the scores vary significantly from items to item? (inter-item reliability); if different people use the same test to rate someone's ability, will they provide the same or very different scores? (Inter-rater reliability). And if an assessment is given over and over again, to what degree do the scores vary across each administration? (Test-retest reliability). Assessment tools should present consistent and coherent pictures of the underlying construct that the aim to capture, but there are many sources of measurement error that can result in inconsistencies. At the child level, when students take a test on different days, scores may randomly fluctuate based on interest, attention, fatigue: for

[10] American Educational Research Association, American Psychological Association, and National Council on Measurement in Education, Standards for Educational and Psychological Testing.

example, a student may score lower on a test on a day where she didn't eat breakfast, but score higher after a nutritious meal. At the inter-individual level, scores may randomly vary based on the testing conditions or on the subjectivity of the rater: one classroom observer may rate the teacher as implementing a practice well, while another rater may score the teacher as implementing the practice poorly. As another example, a child may struggle with a memory test in a tent classroom that is overcrowded, but perform better in his own home.

In turn, the amount of measurement error determines the utility of the scores on the assessment. Researchers, for example, have difficulty detecting the impact of a program if the outcome is measured with a high degree of error. Practitioners and policymakers cannot trust that the data accurately assesses children's learning and well-being, preventing informed decision-making and quality improvement.

Statistics to Provide Evidence on Reliability

- Cronbach's alpha is a widely used metric to assess how well items "hang together", or the internal consistency, of a tool
- Standard error of measurement (SEM) is an item response theory (IRT)-based metric of the precision of a score.
- Kappa, intraclass, and G coefficients can be used to quantify the degree of or variation in agreement among observers.

Validity

Validity refers to the extent to which evidence and theory support the interpretation of scores on an assessment for a specific use: Does the tool measure what it is supposed to measure? There are many types of evidence that can be used to support such interpretations, including evidence of: (1) the test content; (b) the internal structure of the assessment; (c) relationships to other variables; and (d) of the generalizability of the assessment.

In crisis contexts, researchers and practitioners commonly use measurement tools developed for use in WEIRD contexts with little adaptation, raising critical concerns over whether the content of the WEIRD assessment is generalizable: Does it still capture the construct in a different context? For example, if emotional distress in a certain crisis-affected context primarily manifests via somatic symptoms, but the assessment used to measure emotional distress includes questions about psychological symptoms only, evidence about the test content suggests that scores on the assessment are not valid reflections of emotional distress in this context. In turn, this evidence limits researchers' ability to make inferences about whether a certain program impacts emotional distress.

If, however, the scores on the assessment correlate with functional impairment, such "test-criterion" evidence suggests that assessment scores may be able to provide some information – depending on the accuracy of the prediction – about who is having trouble carrying out daily responsibilities. This evidence indicates that practitioners may be able to validly use these scores to screen for participants for a psychosocial life skills program. Given that the validity of the interpretation of the

assessment score depends on the intended use of the assessment, tool developers and users must be clear from the outset about the purpose of the assessment.

Other threats to the valid interpretation of scores on an assessment include systematic factors not related to the construct the assessment is intended to assess. For example, a reading comprehension test administered in a crisis context may contain unfamiliar content (e.g., references to foods or animals not commonly found in that contexts), items that are too hard for the student, or poorly translated words, introducing systematic bias into scores that limit the validity of inferences about scores. In turn, being able to make valid inferences about scores is critical: Otherwise, researchers cannot discern whether programs have meaningful impacts, while practitioners and policymakers cannot trust that the data provides a relevant picture of students' learning and well-being, preventing informed decision-making and quality improvement. Moreover, should stakeholders persist in making such inferences, communities may feel unheard and become disinvested in research and programs.

Methods to Provide Evidence on Validity

- Factor analysis is a statistical technique that can be used to provide evidence that the items on the tool are measuring what they are supposed to measure. It allows one to examine the relationship between observed items or indicators and a smaller number of unobserved latent constructs or factors that are hypothesized to underlie the associations between items.
- Descriptive, correlational analyses can be used to provide evidence of the extent to which:
  - Two measurement tools that assess skills that should be related, are in fact related (convergent validity)
  - A measurement tool predicts a hypothesized outcome at a future time (predictive validity)
  - A measurement tool predicts a hypothesized outcome at the same time point (concurrent validity)
- Review process by a pool of experts can be used to provide evidence of the extent to which the content of a measurement tool -- the themes, dimensions, wording, and format of the items or questions on the tool -- represents the construct it is intended to measure.

Comparability

Comparability refers to the extent to which scores on an assessment can be meaningfully compared across different groups within a context, or across different contexts. There are many factors that may limit the ability to make meaningful comparisons. For example, certain items in an assessment of emotional expression may function differently for girls as opposed to boys, depending on cultural norms around expressing and reporting emotions. As discussed above, too, the construct being assessed may not have the same meaning in different contexts: Emotional distress may mean, and be experienced, quite differently in WEIRD contexts than among former child soldiers in Uganda, for example. Scores on an assessment that does not have evidence of its measurement equivalence – whether a measure is

interpreted and responded to in the same way across two groups – should not be used to make comparisons about levels of emotional distress across contexts. This presents challenges for organizations working in multiple crisis contexts who wish to compare the impact of their programming across contexts, as well as for donors and policymakers who wish to understand how teachers and students in their programs are faring comparatively.

Methods to Provide Evidence on Comparability

- Measurement invariance techniques can be used to provide evidence of the extent to which scores on assessments can be compared across contexts.
- Small-sample qualitative studies or professional judgements can be used to provide explanations of responses to the measurement tool.

Feasibility

Feasibility is not a psychometric criterion, per se, but it is an important functional criterion: Can a measure be successfully administered with limited time, resources, and expertise? Measures that are complex, long, and/or require high levels of technical expertise may require significant human, economic, and physical resources that are not commonly available in crisis contexts. The resulting data may be poor quality, with large amounts of missing data, incorrect scoring, and lacking unique identifying information that allows it be linked to parents', teachers', or peers' data. In turn, this may limit the reliability and validity of scores – and the ability to empirically test the reliability and validity of the scores – while also creating frustration with the data collection process.

For more information, please contact:

Carly Tubbs Dolan
3EA Director of Measurement and Metrics, TIES/NYU
carly.tubbs@nyu.edu

Silvia Diazgranados Ferrans
Senior Research Advisor, Education, IRC
Silvia.diazgranadosferrans@rescue.org

3EA TIES/NYU-IRC DOMAIN, CONSTRUCT, MEASURE, AND LOCATION OF DATA COLLECTION MAP

			Lebanon		Niger		Sierra Leone	DRC	Nigeria	Pakistan	Kurdish Republic of Iraq
Domain	**Construct**	**Measure**	2016-2017	2017-2018	2016-2017	2017-2018	2017-2018	2011-2014	2017-2018	2015-2018	
Academic outcomes	Literacy skills	Annual Status of Education Report	X	X	X	X			X		X
		Early Grade Reading Assessment	X	X	X	X	X	X	X	X	
	Numeracy skills	Annual Status of Education Report	X	X	X	X			X		
		Early Grade Math Assessment	X		X	X		X	X		
Social-emotional well-being and outcomes	Internalizing symptoms	Strengths and Difficulties Questionnaire	X		X			X	X		X
		Moods and Feelings Questionnaire	X	X	X	X			X		
	Externalizing behavior	Strengths and Difficulties Questionnaire	X		X			X	X		X
	Pro-social/ communication skills	Social Competence Scale		X		X			X		
	Classroom social-emotional behaviors	Teacher Observation of Classroom Adaptation				X	X		X		
	Aggression	Children's Stories	X	X	X	X			X		
Social processes	Hostile attribution bias (cognitive information processing)	Children's Stories	X	X	X	X			X		
	Conflict resolution	Children's Stories		X		X			X		
	Social perspective -taking	ACES		X					X		
Emotion processes	Anger dysregulation	Children's Stories	X	X	X	X			X		
	Sadness dysregulation	Children's Stories	X	X	X	X			X		
	Emotion awareness	PANAS-C/P		X							
	Emotion expression	ANAS C/ P+WHO-5		X		X					
	Emotion cue-reading	ACES		X							

			Lebanon		Niger		Sierra Leone	DRC	Nigeria	Pakistan	Kurdish Republic of Iraq
Domain	**Construct**	**Measure**	**2016-2017**	**2017-2018**	**2016-2017**	**2017-2018**	**2017-2018**	**2011-2014**	**2017-2018**	**2015-2018**	
Cognitive processes	Working memory	Rapid Assessment of Cognitive and Emotional Regulation	X	X	X	X					
		Brain Games Executive Functioning Questionnaire	X	X	X	X					
	Inhibitory control	Rapid Assessment of Cognitive and Emotional Regulation	X	X	X	X					
		PSRA - Assessor Report	X	X	X	X					
	Attention/ inattention	Rapid Assessment of Cognitive and Emotional Regulation (Attention Shifting/ Switching)	X	X	X	X					
		Brain Games Executive Functioning Questionnaire	X	X	X	X					
		PSRA - Assessor Report	X	X	X	X			X		
Stress response	Perceived School-related Stress	Response to Stress Questionnaire	X								
	Stress reactivity	Response to Stress Questionnaire	X								
Classroom processes	Classroom climate	Child Friendly Schools Questionnaire	X					X			
	Classroom implementation quality	Teacher Classroom Observation Tool	X	X	X	X	X	X	X	X	X

MEASURES CITATIONS

Measure Name	Acronym	Type of Measure	Adopted, Adapted, Assembled, or Bespoke?	Original Citation(s)	Adapted/Other Citation(s)
Annual Status of Education Report	ASER	Direct assessment	Adapted	Banerji, R., Bhattacharjea, S., & Wadhwa, W. (2013). The Annual Status of *Education Report (ASER). Research in Comparative and International Education*, 8(3), 387–396. https://doi.org/10.2304/rcie.2013.8.3.387	
Early Grade Reading Assessment	EGRA	Direct assessment	Adapted	Gove, A., & Wetterberg, A. (2011). *The Early Grade Reading Assessment: Applications and interventions to improve basic literacy*. RTI International. Retrieved from https://eric.ed.gov/?id=ED531301	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*. Society for Research on Educational Effectiveness. Retrieved from https://eric.ed.gov/?id=ED562783 Kim, H. Y. et al. (2018). *EGRA: Evidence on validity and reliability in Niger and Lebanon.* Unpublished technical report. New York, NY: New York University Global TIES for Children.
Early Grade Math Assessment	EGMA	Direct assessment	Adapted	Gove, A., & Wetterberg, A. (2011). *The Early Grade Math Assessment: Applications and Interventions to improve basic numeracy*. RTI International. Retrieved from https://eric.ed.gov/?id=ED531301	Kim, H. Y. et al. (2018). *EGMA: Evidence on validity and reliability in Niger and Lebanon.* Unpublished technical report. New York, NY: New York University Global TIES for Children.
Strengths and Difficulties Questionnaire	SDQ	Teacher report/ self-report	Adopted with translation	Goodman, R. (1997). The Strengths and Difficulties Questionnaire: A research note. *Journal of Child Psychology and Psychiatry*, 38(5), 581-586.	Tubbs Dolan, C. (2017). *The strengths and difficulties of the Strengths and Difficulties Questionnaire: Cross-national measurement of children's social-emotional well-being in crisis-affected contexts* (Doctoral dissertation). New York University, New York, NY. Diazgranados, S.; Sandoval, A & Carrasco, D (2017). Evidence of Validity and Reliability of 6 Measures of Socio-Emotional Development in Nigeria. Unpublished technical report.

Measure Name	Acronym	Type of Measure	Adopted, Adapted, Assembled, or Bespoke?	Original Citation(s)	Adapted/Other Citation(s)
Moods and Feelings Questionnaire	MFQ	Self-report	Adapted	Messer, S. C., Angold, A., Costello, E. J., & Loeber, R. (1995). Development of a short questionnaire for use in epidemiological studies of depression in children and adolescents: Factor composition and structure across development. *International Journal of Methods in Psychiatric Research*, 5(4), 237-249	Tavitian, L., Atwi, M., Bawab, S., Hariz, N., Zeinoun, P., Khani, M., & Maalouf, F. T. (2014). The Arabic Mood and Feelings Questionnaire: Psychometrics and validity in a clinical sample. *Child Psychiatry & Human Development*, 45(3), 361-368. Kim, H. Y. et al. (2018). *Moods and Feelings Questionnaire: Evidence on validity and reliability in Niger and Lebanon* Unpublished technical report. New York, NY: New York University Global TIES for Children. Diazgranados, S.; Sandoval, A & Carrasco, D (2017). *Evidence of validity and reliability of 6 measures of socio-emotional development in Nigeria*. Unpublished technical report.
Social Competence Scale	SCS	Teacher report	Adapted	Kendall, P. C., & Wilcox, L. E. (1979). Self-control in children: Development of a rating scale. *Journal of Consulting and Clinical Psychology*, 47(6), 1020. Rabiner, D.L., Godwin J., Dodge, K.A. (2016). Predicting academic achievement and attainment: The contribution of early academic skills, attention difficulties, and social competence. *School Psychology Review*. 45(2) 250-267	Kim, H. Y. et al. (2018). *Social Competence Scale: Evidence on validity and reliability in Niger and Lebanon*. Unpublished technical report. New York, NY: New York University Global TIES for Children. Diazgranados, S.; Sandoval, A & Carrasco, D (2017). E*vidence of validity and reliability of 6 measures of socio-emotional development in Nigeria*. Unpublished technical report.
Teacher Observation of Classroom Adaptation	TOCA-C	Teacher report	Adapted	Kellam, S.G., Branch, J.D., Agrawal, K.C., & Ensminger, M.E. (1975). M*ental health and going to school: The Woodlawn Program of assessment, early intervention, and evaluation*. Chicago: University of Chicago Press. Koth, C. W., Bradshaw, C. P., & Leaf, P. J. (2009). Teacher observation of classroom adaptation—checklist: Development and factor structure. M*easurement and Evaluation in Counseling and Development*, 42(1), 15-30.	Kim, H. Y. et al. (2018, in preparation). *TOCA-C: Evidence on validity and reliability in Niger and Lebanon*. Unpublished technical report. New York, NY: NYU Global TIES for Children.

Measure Name	Acronym	Type of Measure	Adopted, Adapted, Assembled, or Bespoke?	Original Citation(s)	Adapted/Other Citation(s)
Children's Stories		Scenario -based report	Adapted and assembled	Dodge, K. A., Price, J. M., Bachorowski, J. A., & Newman, J. P. (1990). Hostile attributional biases in severely aggressive adolescents. *Journal of Abnormal Psychology*, 99(4), 385.	Dodge, K. A., Malone, P. S., Lansford, J. E., Sorbring, E., Skinner, A. T., Tapanya, S., ... & Bacchini, D. (2015). Hostile attributional bias and aggressive behavior in global context. *Proceedings of the National Academy of Sciences*, 112(30), 9310-9315. Di Giunta, L., Iselin, A.-M. R., Eisenberg, N., Pastorelli, C., Gerbino, M., Lansford, J. E., ... Thartori, E. (2017). Measurement invariance and convergent validity of anger and sadness self-regulation among youth from six cultural groups. *Assessment*, 24(4), 484–502. https://doi.org/10.1177/1073191115615214 Kim, H. Y. et al. (2018). *Children's Stories: 3EA adaptation and evidence on reliaiblity and validity in Lebanon and Niger*. Unpublished technical report. New York, NY: NYU Global TIES for Children. Diazgranados, S.; Sandoval, A & Carrasco, D (2017). *Evidence of validity and reliability of 6 measures of socio-emotional development in Nigeria*. Unpublished technical report.
ACES	ACES	Scenario -based report	Adapted	Schultz, D., Izard, C. E., & Bear, G. (2004). Children's emotion processing: Relations to emotionality and aggression.*Development and Psychopathology*, 16(2), 371-387.	Kim, H. Y. et al. (2018). *ACES: 3EA adaptation*. Unpublished technical report. New York, NY: NYU Global TIES for Children.
Positive and Negative Affect Scale + WHO-5	PANAS	Self-report	Assembled	Watson, D., Clark, L. A., & Tellegen, A. (1988). Development and validation of brief measures of positive and negative affect: the PANAS scales. *Journal of Personality and Social Psychology*, 54(6), 1063.	Casuso, L., Gargurevich, R., Van den Noortgate, W., & Van den Bergh, O. (2016). Psychometric properties of the Positive and Negative Affect Scale for Children (PANAS-C) in Peru. *Interamerican Journal of Psychology*, 50(2). Kim, H. Y. et al. (2018). *PANAS: 3EA adaptation*. Unpublished technical report. New York, NY: NYU Global TIES for Children.
Rapid Assessment of Cognitive and Emotional Regulation	RACER	Direct assessment	Bespoke by collaborator	Hamoudi, A., & Sheridan, M. (2015). *Unpacking the black box of cognitive ability: A novel tool for assessment in a population-based survey*. Retrieved from http://thew.web.unc.edu/files/2013/08/hamoudi.pdf	

Measure Name	Acronym	Type of Measure	Adopted, Adapted, Assembled, or Bespoke?	Original Citation(s)	Adapted/Other Citation(s)
Brain Games Executive Functioning Questionnaire		Teacher report	Bespoke by collaborator	Jones, S.M., Bailey, R., and Barnes, S. (2015). *Classroom Executive Function Survey (CEFS): A measure of executive function and self-regulation skills in everyday classroom behavior*. Cambridge, MA: Harvard University.	Kim, H. Y. et al. (2018). *Classroom Executive Functioning Scale: Evidence on reliability and validity in Lebanon and Niger*. Unpublished technical report. New York, NY: NYU Global TIES for Children.
Pre-school Self Regulation Assessment	PSRA	Assessor report	Adapted	Smith-Donald, Raver, Hayes, & Richardson. (2007). Preliminary construct and concurrent validity of the Preschool Self-regulation Assessment (PSRA) for field-based research. *Early Childhood Research Quarterly*, 22(2), 173-187.	McCoy, D. & Raver, C. (2011). Caregiver Emotional Expressiveness, Child Emotion Regulation, and Child Behavior Problems among Head Start Families. Review of Social Development react-text: 58 20(4): 741 - 761 Kim, H. Y. et al. (2018). *PSRA: Evidence on reliability and validity in Lebanon and Niger*. Unpublished technical report. New York, NY: NYU Global TIES for Children.
Response to Stress Questionnaire	RSQ	Self-report	Adapted	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.	Kim, H. Y. et al. (2018). *Response to Stress Questionnaire: Evidence on reliability and validity in Lebanon and Niger*. Unpublished technical report. New York, NY: NYU Global TIES for Children.
Child Friendly Schools Questionnaire	CFS	Self-report	Adapted	Osher, D., Kelly, D. L., Tolani-Brown, N., Shors, L., & Chen, C. S. (2009). *UNICEF child friendly schools programming: Global evaluation final report*. Washington, DC: American Institutes for Research.	Godfrey, E. B., Osher, D., Williams, L. D., Wolf, S., Berg, J. K., Torrente, C., ... Aber, J. L. (2012). Cross-national measurement of school learning environments: Creating indicators for evaluating UNICEF's Child Friendly Schools Initiative. *Children and Youth Services Review*, 34(3), 546–557. https://doi.org/10.1016/j.childyouth.2011.10.015 Kim, H. Y. et al. (2018). *Safe and Supportive Schools Questionnaire: Evidence on reliability and validity in Lebanon and Niger*. Unpublished technical report. New York, NY: NYU Global TIES for Children.
Teacher Classroom Observation	TCO	Observation	Bespoke	Tubbs Dolan, C. (2017). *Improving the quality of education in the Syrian Refugee Response Region: Insights from implementation science* (Doctoral dissertation). New York University, New York, NY.	Diazgranados, S & Linsin, T (2017). *A review of the Teacher Classroom Observation Tool in Lebanon: Evidence of reliabily and convergent and criterion validity*. Unpublished technical report.

COVER PHOTO: PETER BIRO/IRC