



# **Psychometric Analysis of the Pilot Holistic Assessment for Learning (HAL): Validity and Reliability**

Carly Tubbs Dolan

NYU Global TIES for Children

Technical Working Paper

June 2019

## **Abstract**

In this report, we sought to provide evidence on whether data from the pilot version of the HAL instrument could be interpreted as providing consistent and meaningful information about 2<sup>nd</sup> and 3<sup>rd</sup> grade children's literacy, numeracy, and social and emotional skills for formative feedback purposes in the Whole of Syria response. Data were collected from a large sample of Syrian children ( $N = 1,456$ ), the majority of whom were randomly sampled from formal and non-formal education sites in the Northeast, Northwest, and Euphrates Shield regions of Syria. In general, we conclude that there is evidence that scores on the assessment meaningfully capture early grade Arabic literacy and numeracy skills; that scores of literacy and numeracy skills are internally consistent; and that data collectors can provide consistent scores about children's literacy, numeracy, and social-emotional skills. There is limited evidence, however, on the validity of the social and emotional skills subtasks due to issues in the design and implementation of the pilot measure. Recommendations for revision are provided.

---

Correspondence should be directed to Carly Tubbs Dolan at [carly.tubbs@nyu.edu](mailto:carly.tubbs@nyu.edu)

Global TIES for Children, New York University

627 Broadway, Room 807, New York, NY, USA 10012 | [Steinhardt.nyu.edu/ihdsc/global-ties](http://Steinhardt.nyu.edu/ihdsc/global-ties)

## Overview of the HAL: Measurement Library Criteria



The HAL should have moderate to high evidence of validity/reliability for use as a formative evaluation measure. There is promising evidence that a set of subtasks can be used to reliably assess underlying 2<sup>nd</sup> and 3<sup>rd</sup> grade children's early literacy and numeracy competencies; there is also evidence that certain subtasks capture a range of underlying skill levels within the target population. However, given difficulties in the design and administration of the pilot version of the social and emotional skills tasks, there was not conclusive evidence that pilot subtasks reliably captured a set of social and emotional skills as intended. Given that revisions were made to the original version of the HAL based on these pilot findings – and that the pilot was based on a large, representative sample - we preliminarily recommend the literacy and numeracy sections of the revised version of the HAL for formative purposes. However, we urge caution in using the social and emotional skills tasks. We also recommend additional testing of the revised measure.

Criteria	Indicators	Notes
Purpose	Formative	Less consensus on what this requires, but should provide actionable and reliable information about what children know and what they do not yet know.
Empirical evidence overall	# of types of evidence available	4
	% of evidence meets empirical criteria	58% (green only); 74% (green and yellow)
	Evidence fit for purpose	Yes
Confidence in evidence	Sampling method	Stratified random sampling
	Sample size	Large (1,456)
	Missing data	Missingness did not form a critical problem. Missing data, sources of missing data, and how these data were addressed were explicitly reported
	Rigor of method	High
Revisions	Clear guidance on what to adjust/refine	Yes, and revisions were implemented in the version of the HAL that is available for download

## Overview of HAL Empirical Results

Constructs/sub-constructs assessed	Internal structural validity	Internal consistency	Inter-rater reliability	Item functioning	Notes
Early literacy	✓	✓	✓	NA	Recommended revisions are provided on pp. 51-52 of the technical report. These suggestions were taken into account in revising the version of the HAL available for download through the Library.
Letter sound	✓	NA	NA	✓	
Familiar word	✓	NA	NA	✓	
Oral passage reading	NA	NA	NA	NA	
Reading comprehension	NA	NA	NA	NA	
Listening comprehension	NA	NA	NA	NA	
Vocabulary	NA	NA	NA	NA	
Early numeracy	✓	✓	✓	NA	
Number identification	✓	NA	NA	✓	
Number discrimination	○	NA	NA	✓	
Missing number	✗	NA	NA	NA	
Addition	○	NA	NA	✓	
Subtraction	○	NA	NA	✓	
Word problems	NA	NA	NA	NA	
Empathy	✗	NA	NA	NA	
Emotion identification	✗	✗	✓	NA	
Pro-social response	✗	✗	✓	NA	
Hostile attribution bias	✗	✗	✓	NA	
Perspective taking	✗	✗	✓	NA	
Perseverance	○	✓	✓	NA	
Self-concept	○	✓	✓	NA	

### Key

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

This technical working paper was developed by Carly Tubbs Dolan at NYU Global TIES for Children in collaboration with Save the Children and UNICEF, as part of the Whole of Syria Education Cannot Wait initiative. It was reviewed by Jeongmin Lee at the International Rescue Committee for inclusion in the INEE Measurement Library. We thank Clay Westrope (Save), Amy Jo Dowd (formerly at Save), Manuel Cardoso (UNICEF), Reem Khamis-Dakwar (Adelphi University), Karen Froud (Columbia University), Roxane Caires (NYU-TIES) and all participants at the Syria Holistic Assessment regional workshops in August 2018 and April 2019 for their invaluable input and hard work in developing and testing this measure.

**Suggested Citation:** Tubbs Dolan, C. (2019, June). *Psychometric evidence of the pilot Syria Holistic Assessment for Learning: Validity and reliability*. Technical working paper. New York, NY: New York University.

## Table of Contents

Introduction.....	7
The Need for Assessments Developed and Adapted for Crisis Contexts.....	8
The Need for a Holistic Assessment in the WoS Effort .....	10
The Need for Assessments for Formative Purposes in the WoS Effort.....	11
The Need for Contextually Appropriate Assessments.....	13
The HAL Assessment Development Process and Research Aims .....	15
Method .....	16
Sample.....	16
Participants.....	16
Measure.....	17
Literacy. ....	17
Social and Emotional Skills. ....	20
Numeracy. ....	22
Analytic Plan .....	25
Creation of subtask scores. ....	25
RA 1. ....	25
RA 2a, 2b, and 3. ....	26
RA 4. ....	27
RA 5. ....	27
Results.....	28
Literacy .....	28
Descriptive statistics.....	28
Exploratory analyses. ....	31
Confirmatory factor analyses.....	32
Reliability. ....	33
Item difficulty.....	33
Numeracy .....	37

Descriptive statistics.....	37
Exploratory analyses. ....	39
Confirmatory factor analyses.....	40
Reliability.....	41
Item difficulty.....	41
Social and Emotional .....	42
Descriptive statistics.....	42
Exploratory factor analysis.....	47
Confirmatory factor analysis. ....	47
Reliability.....	49
Discussion .....	50
Moving from what children do not know to what children <i>do</i> know .....	50
The importance of contextually appropriate assessments .....	50
Recommendations Based on Psychometric Analyses.....	54
For revisions to test content for formative purposes.....	54
Increase the difficulty level of subtasks with ceiling effects.....	54
Create different version of addition and subtraction subtasks for grades 2 and 3...	54
Replace pictures in empathy subtask with pictures matching the cultural background of children .....	54
For revisions to test administration for formative purposes .....	54
Ensure clarity in stop and skip rule instructions and consistency in application ....	54
Capture information about attempted items.....	54
For scoring for formative purposes.....	54
Literacy. ....	55
Numeracy. ....	55
Social and emotional.....	55
References .....	56

## Introduction

With Sustainable Development Goal (SDG) 4, the global education community promises that all children will have the chance to achieve essential holistic learning and development outcomes as a result of their education. This promise can only be upheld through different types of investments by multiple stakeholders in the most educationally marginalized children: the 75 million children ages 3 to 18 who are currently out of school — and the millions more who are in school but not learning — in crisis contexts (Overseas Development Institute, 2016; UNICEF, 2016a). Importantly, such stakeholders increasingly converge on the importance of investing in the development, adaptation, testing and use of field-feasible, contextually appropriate, and psychometrically sound measurement tools as a strategy for achieving quality education in crisis contexts. These tools can provide accurate and timely data — what Jeffrey Sachs terms the “lifeblood” of the SDGs (Sachs, 2012, p. 210) — about critical dimensions of children’s learning and holistic development to support evidence-based decision-making at different organizational levels (e.g., classroom, district, national) and for various purposes (e.g., formative feedback, program evaluation and monitoring, screening).

The Whole of Syria (WoS)/Education Cannot Wait (ECW) Syria Holistic Assessment for Learning (HAL) tool is one of the first holistic measurement tools designed for use with primary school-aged children to result from such an investment post-2015. The HAL tool – and the process of developing, adapting, and testing the tool – are unique in several ways. First, very few early grade learning assessments have been designed from the outset for use in conflict- and post-conflict settings, which pose unique challenges to the assessment process. Second, the tool is designed to prioritize a holistic approach to learning – capturing social-emotional as well as academic skills -- without compromising the feasibility of assessment. Third, while many initiatives focus on applying rigorous methods to the development of assessments for monitoring and accountability purposes, the WoS/ECW project did so for an assessment designed for formative purposes. The tool is intended to provide teachers and schools in the WoS response region with information about the Arabic literacy, numeracy, and social-emotional skills 2<sup>nd</sup> and 3<sup>rd</sup> grade students have mastered and those that require further support, enabling teachers to identify and implement activities responsive to learning levels. Fourth, the tool was assembled and revised based on consultative process convened by Save the Children and UNICEF to triangulate child developmental research, psychometric, and policy insights from regional literacy and numeracy experts, practitioners with knowledge of the Syrian curriculum, and developmental psychologists with expertise in applied psychometric

analysis. Given often limited time and resources, such an intensive and rigorous measure development and piloting process is rare in crisis contexts.

In this paper, we examine the psychometric properties of the HAL pilot assessment. However, we do so in the context of describing key steps in what we believe is a promising process for developing contextually relevant, field feasible, and psychometrically sound holistic learning measurement tools. In this introduction, we provide a brief overview of challenges to collecting data on learning outcomes in crisis contexts and then introduce the focus and purpose of the HAL assessment, the identification of both of which are critical for developing a tool that provides reliable and valid data fit for purpose. We then briefly describe the HAL consultation and adaptation process, before turning to our research questions.

### **The Need for Assessments Developed and Adapted for Crisis Contexts**

In 2011, the Education for All Global Monitoring Report named four failures – failures to protect, provide, reconstruct, and build peace – by the international education community that have undermined children’s right to education in crisis contexts (UNESCO, 2011). Montjourides (2013) names a fifth: the lack of quality data available in conflict-affected contexts. While many types of data – structural indicators on gross enrollment rates and school retention, spatial and temporal information on the spread and duration of violence – are scarce, particularly pronounced is the lack of coordinated, centralized, and high-quality information on the extent to which children in crisis contexts are learning, a critical target of SDG 4. This includes data on children’s learning outcomes collected from population-based samples that can be used for national and global monitoring purposes, as well as information collected from smaller sub-national or convenience samples for program monitoring and evaluation or formative assessment purposes.

The lack of learning data available for national and global monitoring purposes is in part a reflection of the tremendous human, fiscal, and security resources required for such large-scale efforts. Consider, for example, international assessments of learning outcomes such as Programme d'analyse des systèmes éducatifs de la CONFEMEN (PASEC), the Programme for International Student Assessment (PISA) or the Trends in International Mathematics and Science Study (TIMSS), which will provide the basis for monitoring national progress towards SDG 4.1 targets (Montoya, 2018). Of the 36 countries and territories on the World Bank Group’s fiscal year 2019 harmonized list of fragile situations, nearly two-thirds have never taken part in an international learning



assessment and are not planning to do so in the next cycle of assessment (The World Bank, 2018). Of the 13 countries and territories that either have historically administered such assessments or are planning to participate in upcoming cycles of assessment, three will administer their first assessment in 2019 (the PASEC - Cote d'Ivoire, Democratic Republic of the Congo, and Mali; PASEC, 2017). Three others in the Middle East, however, will not participate in 2019 in assessments they had previously administered (TIMMS – Syrian Arab Republic, Yemen, and West Bank/Gaza; International Association for the Evaluation of Educational Achievement, n.d.).

In terms of smaller-scale data collection efforts, a field scoping study of stakeholders working to promote children's holistic learning and development in crisis contexts undertaken by NYU's Global TIES for Children indicates that holistic learning data is being collected: 90 percent of the 138 researchers and practitioners who responded to that question reported using bespoke and/or established measurement tools to capture information about beneficiaries (Caires & Tubbs Dolan, 2018). However, the most frequently used literacy and social-emotional assessment tools, as reported by some 89 researchers and practitioners, were the Early Grade Reading Assessment (EGRA; 48 percent of respondents; RTI International, 2009) and the Strengths and Difficulties Questionnaire (SDQ; 31 percent of respondents; Goodman, 1997). Both measures were developed based on Western models and assessments of child literacy and social-emotional development, and the extent to which both measures provide meaningful (valid) and accurate (reliable) data in crisis contexts using current scoring and adaptation guidelines has recently been called into question by researchers (Bartlett, Dowd, & Jonason, 2015; Dowd & Bartlett, 2019; Halpin & Torrente, 2014; Tubbs Dolan, 2017). Practitioners agree: 61 percent of practitioner respondents to the scoping study survey "somewhat agreed" or "strongly agreed" that one of the top challenges to regularly monitoring program delivery was the lack of tools they are confident can capture reliable and valid information about program impact and quality.

Under the best of circumstances, establishing the psychometric properties of a measurement tool – its reliability, validity, and comparability – while also ensuring its feasibility of use requires negotiation between scientific rigor and contextual appropriateness. The challenges multiply, however, in crisis contexts: The resource, security, and time constraints that characterize humanitarian settings are such that common methods for meeting one criterion often exacerbate the challenges to achieving others. For example, one common way to improve the internal consistency or reliability of measures is to add items to the measure. While it can prove difficult to lengthen the

data collection time in any context, it becomes particularly unfeasible in crisis-affected contexts for additional reasons, including: (a) security concerns, restricting the amount of time assessors and children can spend in one area; (b) traumatic stress reactions, that limit both children's and assessors' ability to maintain attention; (c) monitoring and evaluation (M&E) funding restrictions, whereby it is seen as unethical to draw additional resources away from urgently needed programming. Negotiating such trade-offs thus requires a clear a priori understanding of the focus and purpose of the measure – as well as where it will be used – in order to best structure and allocate resources for the measurement development and validation process. In the remainder of this introduction, we describe that rationale for the focus and purpose of the WoS/ECW HAL assessment tool, as well as the unique consultative process undertaken to negotiate tradeoffs in feasibility, contextual appropriateness, and scientific rigor.

### **The Need for a Holistic Assessment in the WoS Effort**

Theories from the developmental sciences, decades of research, and practitioners' own experiences working on the ground in humanitarian settings increasingly converge in identifying the need to support both children's academic and social-emotional skill development in crisis contexts (Masten & Narayan, 2012; Aber et al., 2017; Cummings, Merrilees, Taylor, & Mondí, 2017; INEE, 2016; Inter-Agency Standing Committee, 2007). Exposure to armed conflict and its associated adversities – including migration, poverty, and familial separation and loss – pose enduring threats to children's academic (Dryden-Peterson, 2009; Dryden-Peterson, 2011; Kim, Brown, Tubbs Dolan, Sheridan, & Aber, 2019; Trani, Fowler, Bakhshi, & Kumar, 2019) and social-emotional development, including increased risk of mental health disorders (Hadi & Llabre, 1998; Shaw, 2003), hyperactivity (Thabet, Ibraheem, Shivram, Winter, & Vostanis, 2009), aggression (Al-Krenawi & Graham, 2012)), and risk-taking behaviors (Pat-Horenczyk et al., 2007; for reviews in developmental sciences: Cummings, Merrilees, Taylor, & Mondí, 2017; for reviews in public health: Jordans et al., 2016). The limited but growing evidence that has emerged from the Syrian crisis to date indicates that children in Syria and in neighboring host countries face a similar constellation of academic (International Rescue Committee, 2017) and social-emotional difficulties (Khamis, 2019; Perkins, Ajeeb, Fadel, & Saleh, 2018; Sirin & Rogers-Sirin, 2015).

In turn, insults to one developmental domain can bi-directionally and progressively influence other domains of functioning throughout the life course (Masten et al., 2005). For example, Kim et al. (2019) examine how post-migration risk factors at the community, household, and individual level experienced by primary school-aged Syrian

refugee children in Lebanon are associated with cognitive, emotional, and behavioral developmental processes as well as literacy and numeracy performance. They find that children's working memory and behavioral regulation – as assessed by performance-based assessments and observer reports, respectively -- are positively associated with higher literacy and numeracy performance four months later, adjusting for post-migration risk factors. Moreover, executive function and behavioral regulation partially mediated the association between risk factors and academic outcomes. Such holistic quantitative studies are rarely conducted with children in crisis contexts, but they are critical: They suggest the dynamic transactions that occur over time between academic, social, and emotional domains and reinforce the need for a holistic approach to support learning. Moreover, they allow stakeholders to begin to identify the plausible cognitive, social, emotional, and behavioral mechanisms that teachers and caregivers can target and formatively support to promote positive adaptation in a range of settings (Gabrieli, Ansel, & Krachman, 2015; Jones, Greenberg, & Crowley, 2015).

Yet despite increasing recognition of the need for a holistic approach to learning in humanitarian contexts, our review of assessments used with primary school-aged children to date surfaced an either/or approach to assessing skills. That is, they either focus on assessing academic skills such as literacy (e.g., EGRA, RTI International, 2009; Annual Status of Education Review (ASER), Pratham, 2013) and numeracy (EGMA, RTI International, 2009; ASER, Pratham, 2013) or on social-emotional skills, both comprehensive assessments (SDQ, Goodman, 1997; International Social-Emotional Learning Assessment (ISELA), Save the Children, 2019; Developmental Assets Profile - Emergency (DAP-E), Scales et al., 2015; Children and Youth Resilience Measure-12 (CYRM-12), Panter-Brick et al., 2017) and assessments of discrete social-emotional skills (3EA, 2018; Dodge et al., 2015; Ford, Kim, Brown, Aber, & Sheridan, 2019). While such an array of measures can potentially be used in the context of well-funded research studies, it is largely not feasible for teachers or program staff to regularly administer multiple measures – each of which can take 20-30 minutes – during the course of the school day or for routine monitoring and evaluation purposes in crisis contexts. The alternative is the Frankenstein's monster approach: Measures are assembled and subtasks or items cut based on alignment with program goals and contextual relevance. This process largely disregards the psychometric consequences, imperiling the accuracy and meaning – and ultimate utility – of the resulting information.

## **The Need for Assessments for Formative Purposes in the WoS Effort**

Measurement tools provide data that is used for many different purposes – formative, evaluative, screening/diagnosis – at multiple ecological levels (individual, classroom, school, sub-national, national, global). The purpose for and context in which the data will be used then has implications for the design and psychometric properties of the assessment tool. Formative uses – the intended use of the HAL instrument – have two defining characteristics. First, formative assessment by definition involve a feedback loop: Data is interpreted with the goal of deciding how to adjust current practices to be more effective. In the context of a classroom, for example, a teacher may informally ask a question or administer a structured measure in order to learn what curricular content students understand (or don't) and then modify his teaching and learning activities in response. Prompts or items in the assessment should thus provide information that is actionable: What is working? What needs improvement? How can it be improved?

To date, however, international early grade literacy assessments, when used in crisis contexts, have tended to provide more information about what students do not know than what students do know. For example, studies using EGRA subtasks have demonstrated extreme floor effects – where a majority of children score zero – in the Democratic Republic of the Congo (Halpin & Torrente, 2014), Niger (Kim, Brown, Ferrans, & Weiss Yagoda, 2019), Ethiopia (Piper, 2010) and Mali (Spratt, King, & Bulat, 2013). In the Syrian response region, we reviewed two datasets containing information on the literacy skills of Syrian refugee children in Lebanon (N = 2,355; grades 2-4) and Syrian children in Syria (N = 1,467, grade 3), as assessed using different versions of the EGRA and EGMA (see Appendix A, available upon request). While the percentage of students scoring zero on individual subtasks was lower than in other studies cited above, it was not insignificant: In both samples, floor effects were seen in letter sound identification (Lebanon: 31% zero scores; Syria: 18% zero scores), invented word reading (Lebanon: 50% zero scores; Syria: 39% zero scores), oral passage reading (Lebanon: 46% zero scores; Syria: 13% zero scores), and reading comprehension (Lebanon: 54% zero scores; Syria: 31% zero scores) subtasks. When children score zero across a majority of subtasks, and in the absence of other information about the pre-reading skills children do have, it may be particularly challenging for teachers to know how best to scaffold instruction.

Second, assessments used for formative purposes are typically considered lower stakes – and thus needing to meet less stringent psychometric criteria -- than those that are used for evaluation of competencies or diagnosis. Given that such assessments are administered regularly and actions taken based on the data are subject to course correction, the precision of the data is less critical than if it were being used to allocate

resources or opportunities. However, given that the quality of the decision of what to change is in part based on the quality of the data, psychometrically sound formative assessments – when coupled with support and guidance material – may be more effective than those that don’t meet minimum reliability and validity standards.

In the MENA region broadly, the quality of formative assessments has been questioned. For example, the Arab League Educational, Cultural, and Scientific Organization (ALSECO) and the United Nations Education, Scientific, and Cultural Organization (UNESCO) recently undertook a mapping exercise to understand the strengths and weaknesses of assessment systems in Arab countries (UNESCO & ALECSO, 2014).<sup>1</sup> The report concluded that the majority of countries surveyed had in place policy and resource frameworks that provided adequate support for formative, high-stakes, national and international monitoring assessments. However, for formative and national monitoring assessments, the quality of the assessment activity itself – including the psychometric evidence on the instruments and processes and procedures for the assessment – was, on average, weak.

Prior to the civil war, Syria in some ways was a regional leader in prioritization and use of formative assessments: Holistic assessment to feedback information to teachers on both core curriculum skills as well as “non-cognitive skills” was a center piece of the curriculum adopted in 2009 (UNESCO & ALECSO, 2014). Today, as the devastating civil war enters its ninth year, some three million children in Syria do not attend school, in part because some 40 percent of schools are unusable (No Lost Generation, 2019). Nearly 150,000 Syrian teachers have been killed or fled. There are at least seven different curricula being used in Syria as of 2019, and as battle lines have shifted, children have had to jump from one set of learning goals to another (The Economist, 2019). In this context, both the infrastructure for and the quality of assessment activities have been imperiled, making the current effort to develop a contextually appropriate, feasible, and psychometrically sound formative assessment tool linked to support materials and activities even more imperative.

## **The Need for Contextually Appropriate Assessments**

In the past decade a small but growing group of scientists around the world have worked to conceptually and empirically understand child development from within

---

<sup>1</sup> The mapping exercise covered 17 of the 19 Arab countries invited to participate in the study, including: Bahrain, Egypt, Iraq, Jordan, Kuwait, Kingdom of Saudi Arabia, Lebanon, Libya, Mauritania, Oman, Palestine, Qatar, Sudan, Syria, Tunisia, United Arab Emirates, and Yemen.

unique cultural contexts and settings (e.g., Panter-Brick et al., 2017; Oburu & Palmérus, 2016; Yoshikawa & Way, 2008; Eisenberg, Zhou, & Koller, 2001) and to infuse global policy, intervention, and measurement research with culturally appropriate developmental science insights (e.g., Halpin et al., 2019; Jordans et al., 2013; Wuermli et al., 2015; Yoshikawa et al., 2015; Yousafzai, Rasheed, Rizvi, Armstrong, & Bhutta, 2014). Emerging from both strands of research is growing consensus that while there are likely some universal child development processes, how those are manifest, prioritized, defined, and supported varies both across and within cultural contexts and settings. In turn, assessments meant to assess developmental domains such as literacy, numeracy, and social and emotional skills within a certain cultural contexts would likely not provide the most useful or meaningful data unless they have been designed or adapted to capture those manifestations, priorities, definitions, and supports.

For example, the low EGRA scores described above in samples in crisis-affected contexts in the Middle East/North Africa (MENA) and Sub-Saharan Africa (SSA) are likely in part a reflection of the devastating effects that conflict can have on access to and quality of education. However, they may also be an indicator that literacy assessments developed based on Western models of reading are insufficient to meaningfully capture critical language and literacy skills in the complex linguistic contexts in which children in MENA and SSA develop. In the Syrian response region, children learn first to speak a specific dialect of Arabic (e.g., Palestinian Spoken Arabic; notably spoken dialects can vary significantly within and across countries). When they enter formal schooling, however, they are expected to learn a more grammatically complex variant known as Modern Standard Arabic (MSA) for written and formal spoken purposes (Dakwar, 2005). This phenomenon – in which children read a different language than they speak – is known as diglossia, and it is present in Western and non-Western contexts (Ferguson, 1959). Neurolinguistic research has shown that for native Arab speakers, switching between MSA and dialect is neurophysiologically similar to the code-switching processes undertaken in switching from Spanish to English (Khamis-Dakwar & Froud, 2014). This has implications for both test construction and administration. In terms of test construction, for example, it may be easier for children to recognize and pronounce letters and words in MSA – which literacy tests seek to assess – that are close in sound and form to their dialectic counterpart. Tasks that ignore or are inattentive this critical feature of the Arab linguistic and literacy development may be challenging for young readers (Khamis-Dakwar & Khattab, 2014). In terms of test administration, if instructions are not explicit as to whether responses in dialect and/or MSA are allowed -- particularly in non-literacy domains – assessors may introduce additional error introduced into scoring.

### **The HAL Assessment Development Process and Research Aims**

The HAL assessment was designed with the specific assessment content, purpose, and context in mind: It is a holistic assessment intended to provide data for formative purposes to teachers in the Whole of Syria response. This clarity of purpose and form – and knowledge of the necessary implications for test content, structure, and psychometric criteria – guided decision-making throughout a consultative and collaborative test development and piloting process. This consultative process proceeded in five phases: (1) secondary analyses of existing and relevant literacy, numeracy, and social-emotional assessment data in the Syrian response region (see Appendix A, available upon request); (2) triangulation of secondary psychometric results with theory and practice at a workshop with regional experts, curriculum specialists, and developmental psychologists with expertise in applied psychometrics; (3) integration, assembly, and piloting of the HAL; (4) pilot psychometric analyses; and (5) triangulation of psychometric results with theory and practice at a second consultative workshop.

In this paper, we share the results of a set of descriptive and psychometric analyses undertaken with pilot data collected in three provinces of Syria in December 2018 and January 2019. As part of Phase 4, we addressed a set of research aims:

1. To describe the average level of and variation in reading, numeracy, and social-emotional skills among a population-based sample of Syrian children enrolled in grades 2 and 3 in formal and non-formal schools in Syria
2. To provide evidence on the internal structural validity of pilot assessment data, through assessment of: (a) whether the relationships between subtask scores (literacy and numeracy) or item scores (social and emotional) are consistent with a single or multiple underlying constructs; and (b) the strength of relationship of the subtask scores to the underlying construct(s)
3. To test the replicability of the factor structure, through confirming the factor structure in a separate sample
4. To provide evidence on the internal consistency and interrater reliability of the data
5. To assess whether there is evidence that the expected difficulty of items in certain literacy and numeracy subtasks – as hypothesized by the match between MSA and dialect, for example, or curriculum standards -- corresponded to empirical item difficulty.

With the exception of research aim 5, these research aims correspond to those specified in conducting secondary analyses of existing datasets in Phase 1 (see Appendix A, available upon request).

We also conducted a set of exploratory analyses to assess whether there was evidence that the expected difficulty of items in certain literacy and numeracy subtasks – as hypothesized by the match between MSA and dialect, for example, or curriculum standards (see below) -- corresponded to empirical item difficulty. Taken together, the results of these analyses informed but did not determine test revisions to the HAL tool undertaken as part of Phase 5.

## **Method**

### **Sample**

Data from this study come from a sample of Syrian children ( $N = 1,456$ ) in grades 2 and 3, the majority of whom were randomly sampled from formal and non-formal education sites ( $J = 263$ ) in the Northeast ( $n = 259$ ,  $j = 93$ ), Northwest ( $n = 969$ ,  $j = 234$ ), and Euphrates Shield ( $n = 137$ ,  $j = 27$ ) regions of Syria. Sampled sites were 47.9% urban, 42.6% rural, 5.7% mixed, and 3.8% camps for IDPs. On average, five children were sampled per site, although the range varied from two students to fifteen students. An additional 91 children were assessed as enumerators went door-to-door in certain locations given difficulty in locating schools on administrative lists. These home administrations comprised 5% of children sampled in Northwest Syria, 12% of children sampled in Northeast Syria, and 1% of children in Euphrates Shield.

### **Participants**

Children in grade 2 ( $n = 741$ ) were on average 8.05 years old ( $SD = .92$ , range = 6-12) and children in grade 3 were on average 9.15 years old ( $SD = 0.88$ , range = 8-14). On average across the sample, students were 8.6 years old ( $SD = 1.05$ , range = 6-14) and 48% female. The majority of students reported speaking Arabic at home (97.3%), with a small minority reporting speaking Kurdish (1.72%) or multiple languages at home (1%). A majority of children reported living with both their mother and father (88.0%;  $n = 1,281$ ). About half of children reported that their caregiver reads to them and that they see other people reading in the home (54.3%;  $n = 790$ ) but nearly 22 percent of children reported experiencing neither.



## Measure

All children were assessed using the HAL pilot assessment tool. The HAL is organized into three sections corresponding to developmental domain – literacy, social and emotional, and numeracy – which contain anywhere from three to six subtasks, or groupings of items hypothesized to capture information about a specific skill within a developmental domain (e.g., letter sound identification within the literacy domain). Items refer to discrete questions or prompts within subtasks that are scored on a binary (0 = incorrect/no, 1 = correct/yes) scale, with an additional option to record non-response/didn't know (999). HAL subtasks was largely inspired by and assembled from subtasks of existing early learning international assessments (e.g., the Early Grade Reading Assessment, the International Social Emotional Learning Assessment) that – upon extensive review – demonstrated good psychometric properties in use in the Syrian response region, made conceptual sense based on Arabic child developmental theory and research, and aligned with Syrian curriculum goals (see Appendix A, available upon request). Before piloting, however, items within subtasks were revised by a team of regional experts to ensure contextual, developmental, and linguistic appropriateness, and subtasks were further adapted for use as a formative assessment in the Syrian context. The HAL pilot assessment was administered by a trained enumerator to a child in grade 2 or 3 during a break in the school day; the assessment was anticipated to take approximately 30 minutes. Demographic information was also collected from the child at the end of the assessment.

The rest of this section is organized by developmental domain, and it summarizes the types of subtasks and items within each domain; how items and subtasks are scored; and any skip/stop rules applied.

**Literacy.** The literacy domain contains six subtasks designed to assess children's Arabic linguistic processing, decoding, and comprehension skills, as well as their knowledge of Arabic language structure (see Table 1). Subtasks were designed and adapted to account for diglossia and other unique characteristics of the Arabic language (e.g., multiple letter forms, orthography). For each subtask, individual items were summed to create

***Metalinguistic awareness.*** This subtask contains 4 items that ask the child to report on and demonstrate their understanding of the difference between Modern Standard Arabic (fusha) and dialect (ammiya). Skip instructions were not specified in the assessors' guide. However, based on the responses in the dataset, it appears that if

children answered “no” or “don’t know” to questions one (“Do you know that in Arabic there is a fusha and ammiya?”) or three (“Do you at time feel it is hard for you to understand fusha?”), they were not asked questions 2 (“Can you give me an example of when you use fusha and when you use ammiya?”) or 4 (“If you do feel it is hard, is it harder for you when you hear it or when you read it?”), respectively.

***Letter sound recognition.*** This subtask consists of 50 isolated Arabic letter forms presented in a 5 x 10 table intended to assess children’s ability to recognize the letter form and produce – with the guidance of a diacritic -- the corresponding sound in fusha. Letter sounds were ordered within and then across rows by difficulty, as hypothesized by the frequency of the sound and the degree of overlap between the sound in fusha and Syrian dialect. If a child struggled for five seconds with a letter sound, she was prompted with follow-up questions. If she still hesitated, the subtask administration stopped, all unattempted letters were marked as incorrect, and the child proceeded to the next subtask (familiar word reading).

***Familiar word reading.*** This subtask consists of 25 words intended to capture children’s ability to recognize and say in fusha words frequently encountered in the Syrian curriculum. Separate lists of words were developed for grades two and three, and words were ordered by difficulty, as hypothesized by the frequency of the word in Syrian curricular materials and by the degree of overlap in the sound of the word between fusha and Syrian dialect. In test administration, if a child struggled for five seconds with a word, he was prompted with follow-up questions. If he still hesitated, the subtask administration stopped, all unattempted words were marked as incorrect, and the child proceeded to the next subtask (oral passage reading).

***Oral passage reading.*** In this subtask, the assessor asks the child to read a passage of varying length (grade 2 = 82 words; grade 3 = 128 words) in fusha. While the subtask was not timed, per se, the child’s word position at 60 seconds was recorded, potentially enabling calculation of both accuracy and fluency scores. However, in verifying the pilot data, two issues emerged with stop and scoring rules that likely increased the error included in such scores. First, test administration guidance instructed the enumerator to prompt the child to continue if she hesitated on a word; if the child hesitated again, the enumerator was instructed to stop the child, mark the word on which the child stopped and all remaining items incorrect, and move on to the reading comprehension subtask. Due to a programming issue, however, items remaining after the child stopped were marked as correct. In grade 2, the word on which the child stopped was recorded, allowing recoding of skipped items to incorrect/missing. In grade 3, the stop word was

not recorded, so it was not possible to discern which responses were truly correct or incorrect and to calculate a total accuracy score. Instead, for grade 3 I used children's word position at 60 seconds to calculate a total score of words read correctly in 60 seconds, assuming that all prior words were correctly coded as correct/incorrect.

Second, test administration guidance prompted the enumerator to mark a child as a “non-reader” if the child was unable to read five words correctly in thirty seconds. The child was not given the reading comprehension subtask and was prompted to move to the listening comprehension subtask. Due to a separate programming issue, all attempted answers by children marked as “non-readers” were overwritten to missing, resulting in no available information about those children for either the oral passage reading or the reading comprehension tasks. Given these issues, we interpret data and results on this subtask with caution.

***Reading comprehension.*** This subtask consists of four literal and two inferential questions intended to capture children's understanding of the passage they just read aloud. Questions differed by grade level. Test administration guidance instructed the enumerator to prompt the child to continue if the child struggled for five seconds with the question; if the child hesitated again, subtask administration stopped, all questions not attempted were marked as incorrect, and the child proceeded to the next subtask (listening comprehension).

***Listening comprehension.*** The listening comprehensions subtask involves children listening to a pre-recorded passage in fusha and responding to six questions designed to assess their linguistic processing and comprehension skill. No administration guidance was provided for this subtask around stopping rules.

***Expressive vocabulary.*** This subtask requires the child to provide a definition of six words; for students in grade 2, these are words they encountered in the oral passage subtask. No administration guidance was provided for this subtask around stopping rules.

Table 1. *Summary of HAL Literacy Subtasks*

Sub-task	Variable prefix	N of items	Description	Different grade versions ?
----------	-----------------	------------	-------------	----------------------------

Meta-cognition	meta	4	Awareness and demonstration of the difference between fusha and ammiya	No
Letter sounds	lsnd	50	Identification of letter sounds with diacritics, ordered based on match between fusha and ammiya	No
Familiar words	fam	25	Read high-frequency words from Syrian curriculum, ordered based on match between fusha and ammiya	Yes
Oral passage reading	oprfr	82 (G2) 128 (G3)	Read a short passage out loud in fusha	Yes
Reading comprehension	rdcp	6	4 literal, 2 inferential questions about the reading passage	Yes
Listening comprehension	lscpr	6	Listen to re-recorded passage in fusha and answer questions	No
Vocab	voc	6	Expressive vocabulary checking understanding of words from reading passage	No

**Social and Emotional Skills.** The social and emotional skill domain consists of three hypothesized subtasks designed to assess children’s empathy, perseverance, and self-concept (see Table 2). Focal skills were selected in part to align with UNICEF’s Middle East/North Africa Life Skills and Citizenship Engagement framework, and specific subtasks were selected based on the format of assessment (i.e., performance and scenario-based, as opposed to self-report survey) as well as the strength of the (limited) psychometric evidence on social-emotional assessments used previously in the Syrian response region (see Appendix A, available upon request).

***Empathy.*** The empathy subtask consists of three pairs of pictures and short vignettes designed to assess children’s emotion identification, pro-social response, hostile attribution, and perspective-taking skills. Each picture is of a young child expressing a certain negative emotion: sadness, anger, and worry. The assessed child is presented with the first picture and asked to identify the emotion of the child shown in the picture (emotion identification) as well as two things they would do to make the pictured child feel better (pro-social response). The enumerator then reads the child being assessed a short vignette describing the social situation that resulted in the emotion of the pictured child; the social situation is designed to be ambiguous, with a “protagonist” child acting in a way (e.g., spilling juice, stepping ahead in line) towards the pictured child that could be interpreted as hostile in intent. The assessed child is then asked why the protagonist acted that way (hostile attribution bias) and what emotion the protagonist

is feeling (perspective taking). During training, enumerators brainstormed and identified appropriate and inappropriate responses for each item in order to live code during assessment administration correct and incorrect responses.

We note three important issues with the design and administration of this subtask. First, the picture format varies: Sadness was depicted through a drawing while anger and worry were represented by a photograph. Second, all subjects were Caucasian. Given research indicating that the match between the cultural background of the individual expressing an emotion in a still photograph and the judge is important for emotion identification accuracy (see, e.g., meta-analysis by Elfenbein and Ambady, 2002), responses to all items in this subtask could be biased and contain additional error due to format variation. Third, stop rules specified that if a child did not answer the first question (emotion identification) within each picture/vignette pair correctly, enumerators were supposed to mark responses to the other four items in that pair as missing and move to the next picture/vignette pair. Based on a review of the pilot data, however, there is evidence that these rules were inconsistently followed, introducing additional error into the data. Finally, the two pro-social response items within each picture/vignette pair formed a Guttman scale, which implies structural zeros in the two-way tables among items. For example, if a child cannot answer one thing he would do to make the sad child feel better ( $\text{emos\_2}=0$ ), you can assume the child cannot answer a second thing he would do to make the sad child feel better ( $\text{emos\_3}=0$ ). This leaves the  $\{0,1\}$ -cell -- in which a child cannot identify a first pro-social response but can identify a second pro-social response -- necessarily empty. Because this presents computational challenges, binary N Guttman items are recoded into a single item with N+1 response categories. For this analyses, new sadness, angry, and worry pro-social response items -- each with a new scale of 0 – 2 -- were created based on recoding.

***Perseverance.*** In this subtask, children are asked to draw a series of 4 shapes of increasing difficulty with their non-dominant hand. This task is meant to capture children's ability to stay on task despite the task being difficult; however, scores on this task likely contain other sources of variance as well (e.g., familiarity with holding a drawing instrument). If a child completed drawing a shape or was still trying to draw at the one-minute mark, test administration guidance instructed enumerators to mark that item as correct and move on to the next shape. If a child gave up on a drawing within 60 seconds, enumerators were supposed to mark that and further items as incorrect, moving on to the next subtask (self-concept). However, as with the empathy sub-task, in practice it appears enumerators did not consistently follow this stop rule.

**Self-concept.** In this subtask, children are asked to imagine something they hope will happen in the future and respond to questions asking them to describe their future self and identify a barrier to and a key support for achieving this self. Children are then asked to imagine a second thing they hope will happen in the future, and answer the same set of questions. If a child was able to identify any possible future, they were coded as providing a correct response to that item; if not, all subsequent responses were supposed to be marked as missing. However, as with the perseverance and empathy subtasks, in practice it appears enumerators did not consistently follow this stop rule. As such, items in this subtask were treated as a series of plausible Guttman scales (e.g., if a child can't identify one barrier to his future self, it's implausible he would identify a second barrier). For this analysis, items were recoded to form three new items representing the number of future selves, barriers, and supports identified using a zero to two scale. We note, however, that scoring on this subtask should be reconsidered if stop rules are consistently applied.

Table 2. *Summary of HAL SEL subtasks*

Variable	Variable prefix	N of items	Description
Empathy	Emp	15 items across 3 vignettes; 5 items per vignette	For each vignette/picture (sad, angry, worry), child is asked to respond to items on emotion recognition, hostile attribution bias, perspective-taking, pro-social response
Perseverance	Per	4	Child is asked to draw a series of 4 shapes of increasing difficulty with non-dominant hand
Self-concept	Self	6	Child is asked to imagine something they hope that will happen in the future and respond to questions about their hoped-for self (2 items), their concerns for their hoped-for-self (2 items), and agency in achieving their hoped-for-self (2 items)

**Numeracy.** The numeracy domain contains six subtasks designed to assess children's number and operations proficiency with Eastern Arabic/Indian numerals (see Table 3). Subtasks were designed and adapted to account for when and how numeracy concepts are introduced and tested in the Syrian curriculum and to provide procedural information important for formative purposes.

**Number identification.** This subtask contains 20 items designed to assess children's ability to recognize and name Eastern Arabic numerals, ordered by difficulty based on

the number of digits in the number. Test administration guidance instructed the enumerator to prompt the child if the child struggled for 10 seconds on a number; if the child hesitated again, subtask administration stopped, all questions not attempted were marked as incorrect, and the child proceeded to the next subtask (number discrimination).

***Number discrimination.*** This subtask consists of 10 items intended to assess children's ability to make judgements about differences in quantity by comparing sets of numbers. Items included comparisons of one-, two-, and three-digit numbers, ordered by difficulty based on number of digits, distance between numbers within a set (larger differences being easier to discriminate), quantity of unit digit (higher unit digits being harder to discriminate), and compatibility of comparisons (a two- or three-digit number in which all digits are smaller or larger than the digits in the other number in the set being easier to discriminate). Test administration guidance instructed the enumerator not to prompt the child after one example; if the child hesitated for 10 seconds at any point, that item was marked as incorrect and the child proceeded to the next item until all 10 items were completed. The child then moved on to the next sub-task (missing number).

***Missing number.*** In this task, children are asked to identify the missing number in a sequence of four numbers. Number sequences are ordered by hypothesized difficulty, and sequences included variation in the number of digits (1, 2, 3) of numbers in a sequence; position of the missing number; distance between numbers (1, 2, 5, 10, 100); familiarity of starting number; and forward/backward counting. Test administration guidance instructed the enumerator not to prompt the child after one example; if the child hesitated for 10 seconds at any point, that item was marked as incorrect and the child proceeded to the next item until all 10 items were completed. The child then moved on to the next subtask (addition).

***Addition.*** The seven items in this subtask are designed to assess children's ability to solve one- and two-digit addition problems. Items were ordered by difficulty in the following sequence: (1) 2x1 digit number, first digit smaller; (2) 2x1 digit numbers, first digit larger; (3) 1 digit plus 2 digit; (4) two digit plus one digit; (5) two digit plus two digit, no carry; (6) two digit plus two digit, carry; and (7) two digit plus two digit, first digit smaller. Addition problems were written using Eastern Arabic numerals and were presented to children in either horizontal or vertical format (for problems requiring addition in both the unit and the tens columns). Children were given scratch pen and paper and allowed to use their fingers to solve the problem. If the child hesitated for 30

seconds on any problem, test administration guidance instructed the enumerator to mark all remaining items incorrect and move on to the next subtask (subtraction).

**Subtraction.** The six items in this subtask are intended to assess children’s ability to solve one- and two-digit subtraction problems. Items were ordered by hypothesized difficulty based on the following sequence: (1) 2x1 digit number, first number largest; (2) 2 digits minus 1 digit, no carry; (3) 2 digit minus 1 digit, no carry; (4) 2 digit minus 2 digit, no carry; and (5) 2 digit minus 2 digit, carry. A sixth item was also included that asked children to solve a 2x1 digit addition equation in which one of the addends was left blank. Format and stop rules were the same as for addition.

**Word problems.** The six items in this subtask are designed to assess children’s ability to apply mathematics concepts and operations in familiar situations and to provide information on key problem solving strategies children have mastered (or not). Four items asked children to directly provide a response to addition/subtraction word problems that required children to join/separate quantities, identify the relationship of parts to a whole, and/or compare quantities. Two additional word problems focused on assessing procedural knowledge; each contained two sub-questions to gauge children’s ability to: (1) comprehend the word problem; (2) make a plan to solve the word problem; (3) execute the plan; or (4) verify that the answer is correct. In scoring these two procedural sub-questions, both responses had to be correct in order to receive a “correct” score on that word problem; if only one response was correct, the child received an “incorrect” score on that word problem. Different word problem versions were developed and administered for grades 2 and 3. Children were provided with scratch paper to help in solving the problems, and no stop rules were specified in the administration guidance.

Table 3. *Summary of HAL Numeracy Subtasks*

Variable	Variable prefix	N of items	Description	Different grade versions?
Number identification	nid	20	Child asked to identify numbers, ordered by difficulty	
Number discrimination	dis	10	Child asked to identify larger number between two numbers	
Missing number	miss	10	Child asked to identify missing number in series of numbers with increasing pattern difficulty	



Addition	add	7	Child asked to add single and double digit numbers, ordered by difficulty	
Subtraction	sub	6	Child asked to subtract single and double digit numbers ordered by difficulty; 1 item that is “fill in the blank” operator	
Word problems	word	6	Word problems including <ul style="list-style-type: none"> <li>• Traditional approach: simple solution of problem</li> <li>• New approach: explain comprehension, plan, execution, and/or verification</li> </ul>	Yes

## Analytic Plan

**Creation of subtask scores.** For both descriptive and analytic purposes, literacy and numeracy subtask scores were created by taking the unweighted sum of items comprising individual subtasks. This subtask score corresponds to the total number of items answered correctly on the subtask, and it treats items not attempted because of stop/skip rules as incorrect (zero-imputed).<sup>2</sup> Given that oral passage reading was a timed subtask, a second sum score corresponding to the number of items answered correctly on the subtask in one minute was created. Given the highly exploratory nature of analyses with the social-emotional items, subtask scores were not created a priori.

**RA 1.** In order to understand the range and nature of variation in Syrian children’s literacy, numeracy, and social-emotional skills as assessed using the HAL, a set of descriptive statistics – means, standard deviations, range – were calculated, and histograms showing distributional characteristics were created. In order to assess the associations between subtask scores within and across dimensions, I computed bivariate correlations between literacy, numeracy, and social-emotional sub-task scores.

---

<sup>2</sup> Such zero-imputed sum scores make two important assumptions: (1) A child would have gotten all items wrong after stopping; (2) Non-readers scores on oprf and rdcp are equivalent to “readers” who attempted and answered items incorrectly. To test these assumptions I also calculated a version of subtask scores in which items not attempted because of stop/skip rules were treated as missing, and conducted sensitivity analyses for Aims 2 and 3. Results were largely similar to those reported herein and are available upon request.

**RA 2a, 2b, and 3.** In order to provide evidence of the internal structural validity of the HAL assessment, I conducted a series of exploratory and confirmatory factor analyses within each developmental domain using the following steps:

***Step 1: Randomly split sample.*** I first randomly divided the full sample in half stratified on gender and grade in order to create exploratory and confirmatory samples. 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 hypothesized and proposed factor structures, thereby builds confidence in the stability of empirically derived exploratory factor analytic estimates (Osborn & Fitzpatrick, 2012).

***Step 2: Decide on level of test analysis by domain.*** The analytic approach undertaken within the literacy/numeracy and social-emotional domains to address research aims 2a and 2b differed in the level of aggregation of assessment data. Given the limited number of social-emotional items and subtasks, a series of exploratory and confirmatory factor analyses were conducted with the individual items scores within the domain. For literacy and numeracy, analyses were conducted with subtask scores within a domain. This approach – also taken in conducting secondary analyses of other early grade learning assessments in Phase 1 (see Appendix A, available upon request) – is similar to that used by Halpin, Torrente, & Aber (2016) to assess the internal structure of the EGRA in the DRC. This approach was originally selected and applied to the secondary analysis datasets during Phase 1 given that the number of zero scores on individual items within each subtask created challenges to model convergence. We apply the same approach to HAL literacy and numeracy data to ensure alignment with the secondary analyses and given that for formative purposes teachers will likely aggregate data to the subtask level.

However, this approach makes the assumption that all items within a subtask discriminate to the same degree between underlying ability levels (within an item response theory (IRT) framework) or load equally onto a latent subtask construct (within a factor analytic framework). While making such an assumption is generally considered acceptable for exploratory purposes such as this (DiStefano & Zhu, 2009; Tabachnick, Fidell, & Ullman, 2007), we note that all subtask-level factor analytic results should be interpreted in the context of this assumption.

***Step 3: Conduct exploratory and confirmatory analyses within domain.*** In order to account for structural characteristics of the data, two important specifications

are made. First, item and subtask response distributions followed many different functional forms. Because modeling the responses as normally and continuously distributed can lead to inflation of model fit statistics and biased estimation of factor loadings and standard errors, I used a weighted least squares mean and variance-adjusted (WLSMV) estimator and specified whether the functional form was categorical, censored, or continuous (see Appendix Table 1; Beauducel & Herzberg, 2006; Lei, 2009). Second, as described above, many of the students in this sample were nested in education sites, thereby violating the assumption of independence of standard errors required in the application of factor analytic techniques. I thus used the TYPE=COMPLEX command in MPlus v8.0 to estimate robust standard errors (Muthén & Muthén, 2012).

Within each domain, I then fit a series of exploratory factor analysis models using an oblimin rotation (Marsh, Morin, Parker, & Kaur, 2014). To assess the goodness of fit of the models, the following two criteria were used (Hu & Bentler, 1999): (a) a root mean squared error of approximation (RMSEA) value below .08 provides an acceptable fit to the data, while an RMSEA of less than .05 provides a good fit to the data; and (b) a comparative fit index (CFI) value above .9 provides an acceptable fit to the data while a CFI value above .95 provides a good fit to the data (Kline, 2011). Based on a combination of the overall goodness-of-fit statistics, item specificity and strength, visual inspection of the residual correlation matrix, and face validity of the models, I then selected a factor structure to test for each domain using the confirmatory sample.

**RA 4.** In order to provide evidence of the internal consistency of the HAL, I calculated Cronbach's alpha for each empirically derived domain scale (Cronbach & Shavelson, 2004). In order to provide evidence of the interrater reliability of the HAL, I calculated Krippendorff's alpha weighted for ordinal data for each rater pair, and averaged across rater pairs to arrive at an estimate of IRR for subtasks (literacy/numeracy) and items (social-emotional). Although less frequently encountered in the social, behavioral, and education sciences, Krippendorff's alpha – a measure of the proportion of agreement among raters above what would be encountered by chance -- allows for greater flexibility in the number of raters, degree of missing data, and scale of data than more common statistics (Hayes & Krippendorff, 2007; Krippendorff, 2011; Stemler & Tsai, 2008).

**RA 5.** Finally, in order to assess whether there is evidence that the expected difficulty of items in certain literacy and numeracy subtasks corresponded to empirical item difficulty, I fit a series of full information maximum likelihood one- and two-parameter

(2-PL) models to dichotomous item response data within letter sound, familiar word, number identification, number discrimination, missing number, addition, and subtraction subtasks (Lord, 2012). Models were estimated with a quadrature (EM) algorithm approach using rectangular Monte Carlo grids (Bock & Aitken, 1981).

Broadly, this family of models relate examinees' underlying ability ( $\theta$ ) and item parameters using logistic functions. In order to ensure the assumptions of unidimensionality required for use of such models were met, I first fit a unidimensional confirmatory factor model to the item response data and evaluated the fit according to the criteria described above (Hambleton, Swaminathan, & Rogers, 1991). I then compared the fit of the 1- and 2-PL models and proceeded with item analysis using the better fitting model. In examining items, I focused on the b-parameter, or difficulty parameter (Embretson, Reise, & Reise, 2013). The b-parameter is at the point on the  $\theta$  scale where the probability of a correct response is equal to 0.50 and typically varies from -2.00 to 2.00, increasing as items become more difficult (Embretson et al., 2013; Hambleton et al., 1991).

## Results

Results are reported by developmental domain and within developmental domain, by research aim.

### Literacy

Given that different versions of familiar word reading, oral passage reading, and reading comprehension subtasks were administered for each grade, analyses were conducted separately for each grade.

**Descriptive statistics.** Table 4 contains descriptive statistics for subtask total scores – the total number of items answered correctly for each subtask -- by grade, while Table 5 displays the pairwise correlation matrix among subtask scores.

Table 4. *Literacy subtask descriptive statistics*

Variable	n <sub>g2</sub>	Grade 2 mean/%	Grade 2 SD	n <sub>g3</sub>	Grade 3 mean/%	Grade 3 SD	Min	Max
meta_1	741	28.07%		715	49.09%		0	1
meta_2	208	38.94%		351	49.57%		0	1
meta_3	741	25.10%		715	32.73%		0	1

meta_4	186	69.89%		234	60.68%		0	1
lsnd_t	741	39.98	14.12	715	44.22	11.44	0	50
fam_t*	741	17.99	7.45	715	19.74	7.08	0	25
opr_f_t**	741	43.52	32.93		NA	NA	0	81
opr_f_mn*	741	19.74	16.85	713	40.36	26.82	0	76/126
rdcp_t*	741	2.88	2.46	714	3.94	2.09	0	6
lscp_t	741	4.41	1.49	715	5.07	1.36	0	6
voc_t	741	4.06	1.60	715	4.62	1.43	0	6

\*Different versions administered in grades 2 and 3

+ Given data collection challenges, the total score variable could not be calculated for grade 3 students. The subsequent row in the table contains the oral passage reading variable corresponding to the total number of words read correctly in one minute

Table 5. *Pairwise correlations among literacy subtask scores*

	lsnd_t		fam_t		opr_f_t		opr_f_tmn		rdcp_t		lscp_t		voc_t	
lsnd_t	1	1												
fam_t*	0.74	0.75	1	1										
opr_f_t**	0.43	-	0.57	-	1	-								
opr_f_mn*	0.43	0.46	0.55	0.60	0.81	-	1	1						
rdcp_t*	0.45	0.55	0.57	0.74	0.85	-	0.78	0.65	1	1				
lscp_t	0.36	0.42	0.42	0.47	0.41	-	0.37	0.34	0.48	0.52	1	1		
voc_t	0.40	0.52	0.46	0.55	0.50	-	0.46	0.44	0.56	0.57	0.56	0.56	1	1

Note: All correlations significant at the  $p < 0.000$  level

\*Different versions administered in grades 2 and 3

+ Given data collection challenges, the total score variable could not be calculated for grade 3 students. The subsequent row in the table contains correlations among subtasks and the oral passage reading variable corresponding to the total number of words read correctly in one minute.

Figure 1. *Grade 2 literacy subtask distributions*

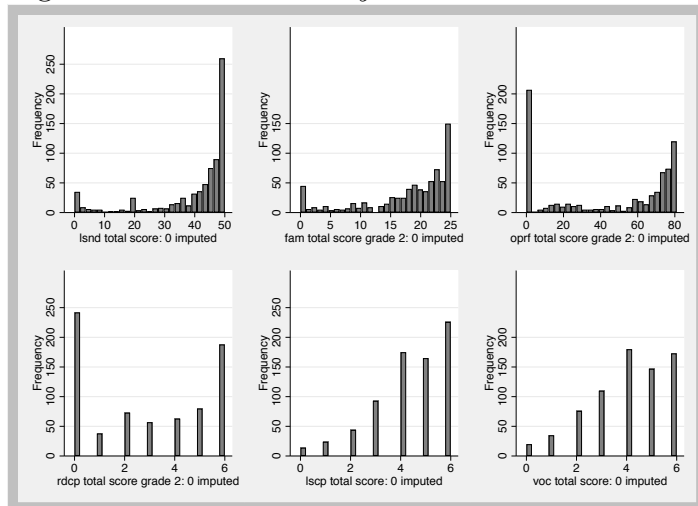
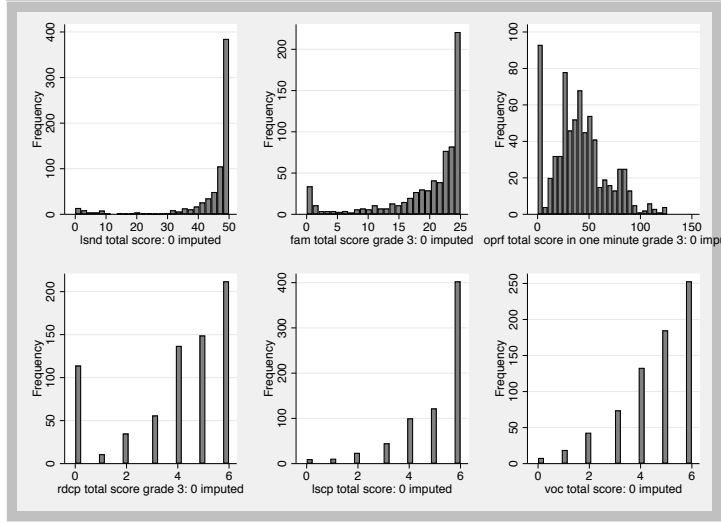


Figure 2. *Grade 3 literacy subtask distribution*



***Diglossic awareness.*** A minority of students in the sample (38%) reported knowing that in Arabic there is a fusha and ammiya. Of those that did, only 46% were able to give a correct example of when to use fusha and ammiya. Nonetheless, only 29% of students reported that it was at times hard for them to understand fusha. Of those that did find it at times hard, more students reported difficulty reading fusha (65%) than hearing fusha (35%).

***Literacy skills.*** Students in both grades 2 and 3 on average scored highly on the letter sound (mean percent correct ( $MPC_{g2}$ ) = 80%,  $MPC_{g3}$  = 88%), familiar word ( $MPC_{g2}$  = 72%,  $MPC_{g3}$  = 79%), listening comprehension ( $MPC_{g2}$  = 73%,  $MPC_{g3}$  = 84%), and expressive vocabulary ( $MPC_{g2}$  = 68%,  $MPC_{g3}$  = 78%) tasks; as expected, for shared subtasks students in grade 3 had higher scores than students in grade 2. All four of these subtasks display ceiling effects, although the negative skewness is more pronounced in grade 3.

Students in grade 2 struggled more with oral passage reading ( $MPC_{g2}$  = 53%) and reading comprehension ( $MPC_{g2}$  = 48%) tasks. Although the average scores are near the scale midpoints, scores are not normally distributed: the distributions show evidence of both floor effects (due primarily to the “non-reader” group; see description under “Measure” sub-section) and ceiling effects. Grade 3 students on average scored higher on oral passage reading ( $MPC_{g3}$  = 79%) and reading comprehension ( $MPC_{g3}$  = 65%) subtasks. The grade 3 oral passage reading distribution approximated a zero-inflated

distribution, in which there is a set of zero scores (due to the non-reader classification) while other scores approach a normal distribution.

***Subgroup differences in literacy skills.*** Children who reported knowing that there was a fusha and ammiya in Arabic scored significantly higher on letter sound ( $t(1454)=10.65$ ,  $p = 0.000$ ), familiar word ( $t_{g2}(739)=5.72$ ,  $p = 0.000$ ;  $t_{g3}(713)=8.98$ ,  $p = 0.000$ ), oral passage reading ( $t_{g2}(739)=-4.72$ ,  $p = 0.000$ ;  $t_{g3}(713)=7.08$ ,  $p = 0.00$ ), reading comprehension ( $t_{g2}(739)=6.96$ ,  $p = 0.000$ ;  $t_{g3}(713)=8.10$ ,  $p = 0.00$ ), listening comprehension ( $t_{g2}(1454)=9.31$ ,  $p = 0.000$ ), and expressive vocabulary subtasks ( $t_{g2}(1454)=11.31$ ,  $p = 0.000$ ).

Students in grades 2 and 3 marked as “non-readers” by their assessors – based on not being able to correctly read at least five words in the oral passage in 30 seconds – scored significantly lower than those marked as readers on letter sound ( $t_{g2}(739)=15.89$ ,  $p = 0.000$ ;  $t_{g3}(713)=22.33$ ,  $p = 0.000$ ), familiar word ( $t_{g2}(739)=20.64$ ,  $p = 0.000$ ;  $t_{g3}(713)=28.67$ ,  $p = 0.000$ ), oral passage reading ( $t_{g2}(739)=38.21$ ,  $p = 0.000$ ;  $t_{g3}(713)=19.15$ ,  $p = 0.00$ ), reading comprehension ( $t_{g2}(739)=28.47$ ,  $p = 0.000$ ;  $t_{g3}(713)=28.53$ ,  $p = 0.00$ ), listening comprehension ( $t_{g2}(739)=9.89$ ,  $p = 0.000$ ;  $t_{g3}(713)=9.84$ ,  $p = 0.00$ ), and expressive vocabulary subtasks ( $t_{g2}(739)=13.56$ ,  $p = 0.000$ ;  $t_{g3}(713)=14.50$ ,  $p = 0.00$ ). However, “non-readers” did not on average score zero on emergent literacy tasks: For example, non-readers in both grades on average correctly responded to 55% of letter sound items, 62% of listening comprehension items, and 48% of vocabulary items.

**Exploratory analyses.** Exploratory factor analyses using all literacy subtasks were conducted separately with the grade 2 and the grade 3 exploratory samples. As shown in Table 6, one- (Model A) and two- (Model B) factor solutions did not provide a good fit to the data in either grade. Based on prior research suggesting that listening comprehension does not load on to an early literacy latent factor (Halpin & Torrente, 2014) as well as the large residual variance for that item in both Models A and B, I removed listening comprehension and refit the one-factor model (Model C). This model also did not provide a good fit to the data. Finally, inspection of the residual correlation matrix of the one-factor model indicated that the largest residual involved the letter sound subtask. I thus tested one-factor models in which I: (i) removed letter sound subtask scores (Model D); and (ii) retained letter sound subtask scores and allowed for a residual correlation between letter sound and familiar word reading (Model E). Model D provided an acceptable fit to the data in grades 2 and 3, while Model E provided a good fit to the data in grade 2 and an acceptable fit in grade 3. This suggests that a one-

factor “early literacy” model is acceptable for describing correlations among familiar word reading, oral passage reading, reading comprehension, and expressive vocabulary subtasks. However, based on the residual correlation between familiar word reading and letter sound identification – as well as the improved model fit when letter sound is removed from the model – it is possible that there may be two distinct but related early literacy constructs in grade 2. Unfortunately, empirically testing a model with a two-item factor is not recommended given difficulties identifying residual covariances between items (Muthén & Muthén, 2014).

Table 6. *Exploratory analysis model fit*

		Model A	Model B	Model C	Model D	Model E
Grade 2	X <sup>2</sup>	90.784*	62.631*	67.28	6.850	5.905
	df	9	4	5	2	4
	RMSEA	0.157 (0.128	0.199 (.157-	0.183 (.146-	0.081 (.020-	0.036 (.00-
	(90% CI)	0.187)	.244)	.224)	.151)	.092)
	CFI	0.884	0.917	0.856	0.989	.997
Grade 3	X <sup>2</sup>	55.01	17.317	21.856	5.628	5.628
	df	9	4	5	2	2
	RMSEA	0.120	0.097	0.097	0.071	0.071
	(90% CI)	(0.090-0.151)	(0.053-0.145)	(0.058-0.141)	( 0.000-0.144)	( 0.00-0.144)
	CFI	.959	0.988	0.980	0.995	0.995

Note: X<sup>2</sup> and df refer to the chi-square test off model fit and its degrees of freedom; RSMEA (90% CI) denotes the root mean square error of approximation and its 90% confidence interval; CFI denotes the comparative fit index.

**Confirmatory factor analyses.** While Model E provided a better fit to the data, the residual correlation between letter sound and familiar word reading subtasks would require the use of refined scoring methods not optimal for a formative assessment. I thus proceeded with confirming Model D using the grade 2 and 3 confirmatory sample. This model provided a good fit to the data in both grades 2 (X<sup>2</sup> = 1.04; RMSEA = 0.00 (0.00-.10); CFI = 1.00) and 3 (X<sup>2</sup> = 1.89; RMSEA = 0.00 (0.00-.10); CFI = 1.00). Table 7 shows the standardized factor loadings for each final confirmatory model. All items load highly on the underlying factor ( $\lambda > .700$ ) with the exception of vocabulary in grade 2. However, we emphasize that comparisons should not be drawn between grades about the magnitude of factor loadings given that each model contains different subtasks due to both planned administration and data collection challenges.

Table 7. Factor loadings for confirmatory models

	Grade 2	Grade 3
fam_t*	.743	.877
oprf_t*+	.845	-



opr_f_mn* <sup>+</sup>	-	.778
rdcp_t*	.995	.877
voc_t	.659	.877

Note: All factors were significant at  $p < .001$ , after adjusting for clustered standard errors.

\*Different versions administered in grades 2 and 3

<sup>+</sup> Given data collection challenges, the total score variable could not be calculated for grade 3 students. Instead, the total number of words read correctly in one minute was used in factor analyses for the grade 3 sample.

## Reliability.

**Internal consistency.** Cronbach's alpha coefficients of literacy scores derived from the final confirmatory model were acceptable ( $\alpha_{g2} = .85$ ,  $\alpha_{g3} = .85$ ).

**Interrater reliability.** Krippendorff's alpha coefficients for each literacy subtask are shown in Table 8. All coefficients are above the common benchmark for substantial interrater agreement ( $\alpha > .80$ ).

Table 8. Krippendorff's alpha coefficients

lsnd_t	fam_t*	opr_f_t* <sup>+</sup>	opr_f_mn*	rdcp_t*	lscp_t	voc_t
.95	.97	.99	.82	.93	.98	.97

## Item difficulty.

**Letter sound.** Confirmatory factor analyses indicated that a one-factor model provided an excellent fit to item-level letter sound subtask data (grade 2:  $X^2(1175) = 2213$ ; RMSEA = 0.35 (0.032-.037); CFI = .979; grade 3:  $X^2(1175) = 1522$ ; RMSEA = 0.20 (0.017-.023); CFI = .992). I then fit 1-PL and 2-PL models to the data and compared the fit: In both grade 2 ( $X^2(49) = 490.10$ ,  $p = 0.00$ ;  $\Delta AIC = -393$ ;  $\Delta AIC = -169.98$ ) and grade 3 ( $X^2(49) = 387.80$ ,  $p = 0.00$ ;  $\Delta AIC = -289.82$ ;  $\Delta AIC = -65.99$ ), the 2-PL model provided a better fit to the data than the 1-PL model.

Figure 3 shows the item difficulty level – the b parameter in the 2-PL model – for each item by grade. There are several things to note about these difficulty levels. First, the negative valence on all parameter estimates indicates that items on this subtask are easy for children whose underlying letter sound skill level is below average. For children in grade 2, test items provide information about children who are -.5 standard deviations (SD) to -2 SD below the mean; for children in grade 3, test items provide information about children who are -1 SD to -2 SD below the mean. Second, all items are easier for grade 3 students than grade 2 students, but they vary in the extent to which they are easier. For example, the underlying ability level at which children are likely to correctly

answer items 3 and 45 does not appear to greatly decrease in grade 3. In contrast, children in grade 3 at a lower latent skill level than in grade 2 are able to correctly answer items 19 and 28, signaling that these items became easier. Third, visual inspection of the empirical item difficulties indicated some convergences and some divergences with the hypothesized ordering of items by difficulty based on match in sound between MSA and dialect (recall that items were ordered by difficulty within groups of five). Most strikingly, item 1 is one of the two easiest items while item 50 is the most difficult item. Items that greatly diverge from the hypothesized ordering are marked in Figure 3 with a black diamond; they constitute roughly 19% of all items.

***Familiar word.*** Confirmatory factor analyses indicated that a one-factor model provided an excellent fit to item-level familiar word subtask data (grade 2:  $X^2(275) = 464.23$ ; RMSEA = 0.30 (0.026-.035); CFI = .987; grade 3:  $X^2(275) = 377.92$ ; RMSEA = 0.23 (0.017-.028); CFI = .99). I then fit 1-PL and 2-PL models to the data and compared the fit: In both grade 2 ( $X^2(24) = 127.45$ ,  $p = 0.00$ ;  $\Delta AIC = -79.45$ ;  $\Delta BIC = 30.84$ ) and grade 3 ( $X^2(24) = 93.27$ ,  $p = 0.00$ ;  $\Delta AIC = -45.27$ ;  $\Delta BIC = 64.19$ ), the 2-PL model provided a better fit to the data than the 1-PL model as indexed by the  $X^2$  and  $\Delta AIC$ .

Figure 3. *Item difficulties (2-PL) for letter sound subtask, by grade*

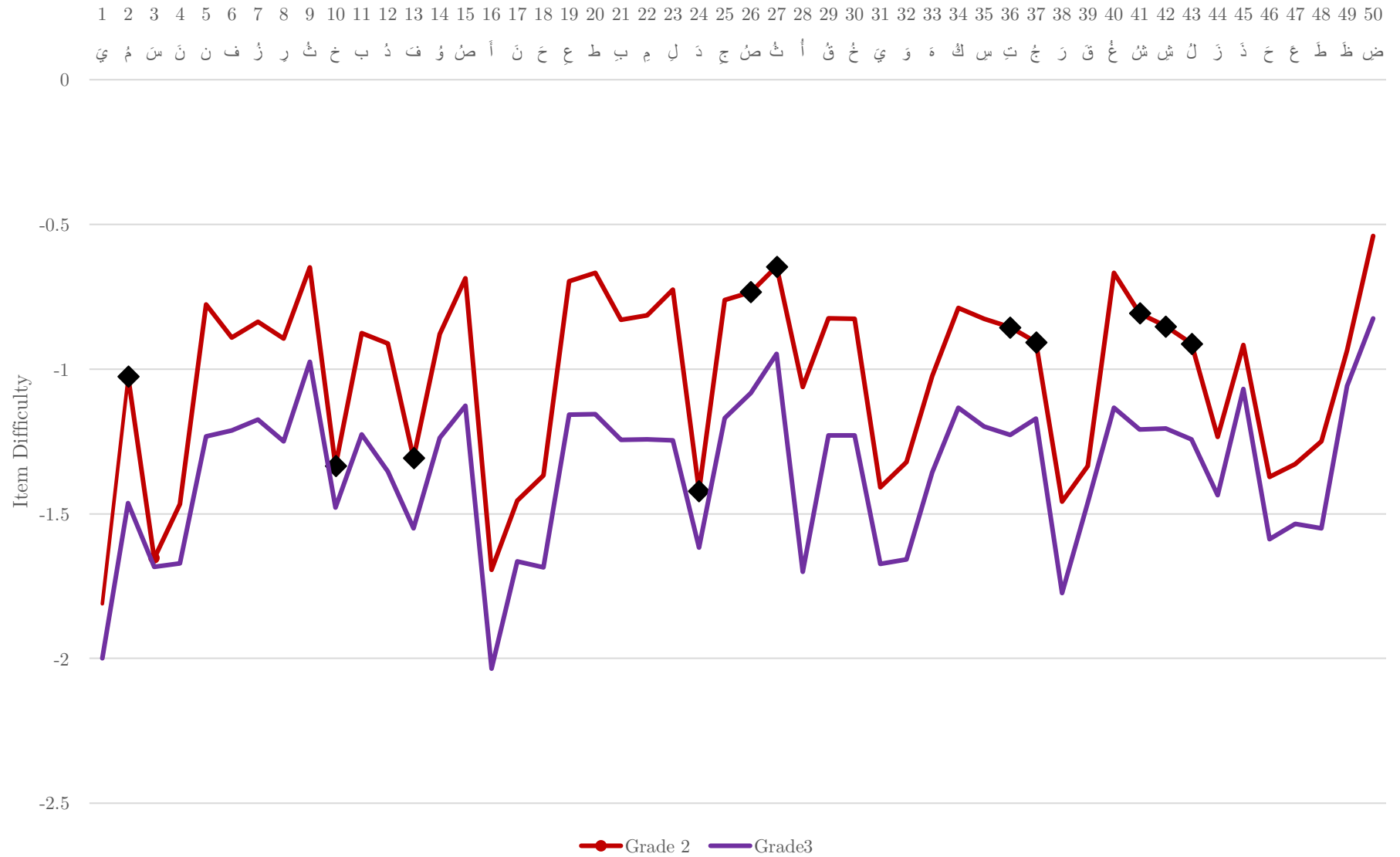
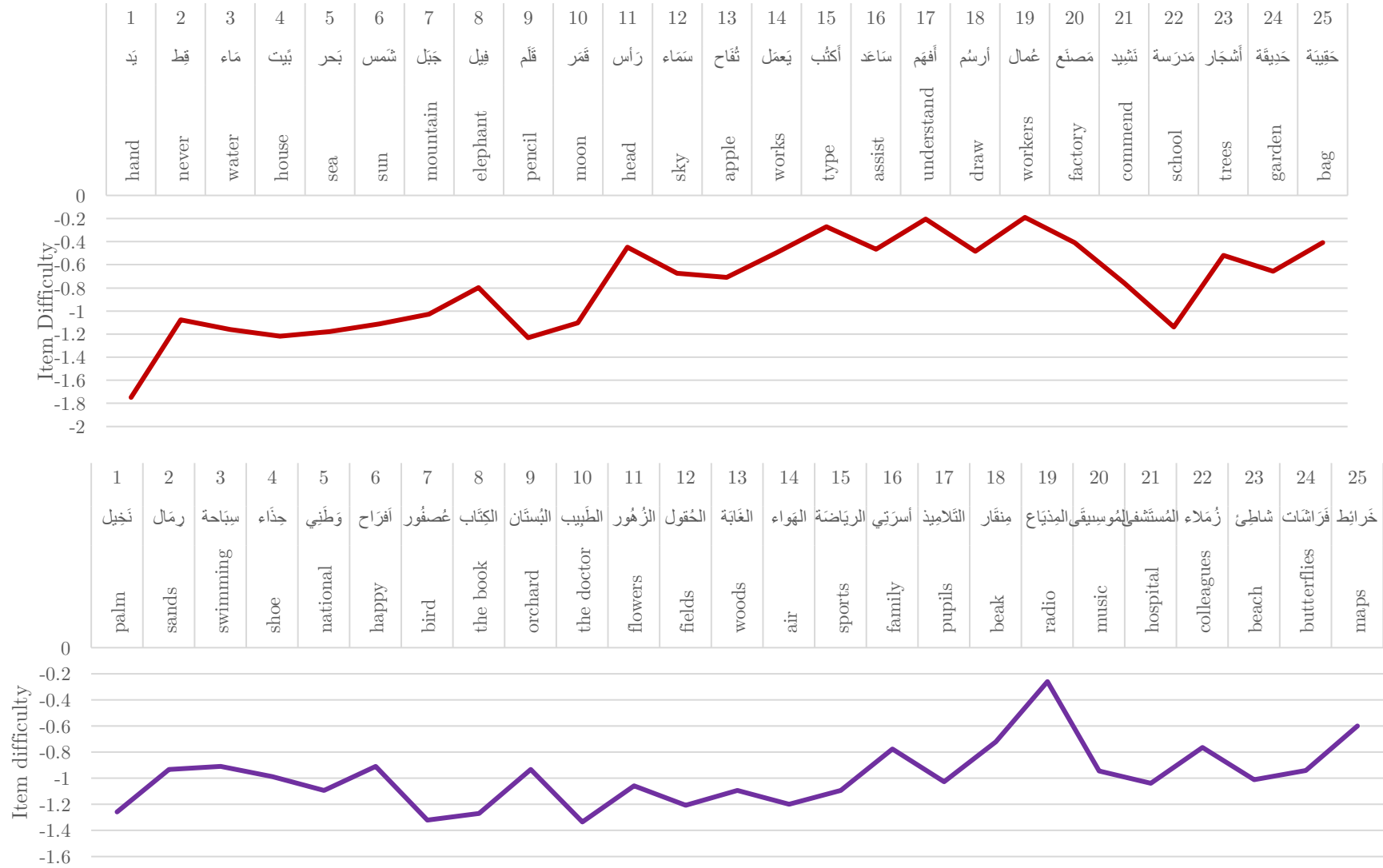


Figure 4. Item difficulties (2-PL) for familiar word subtask, by grade



Note: Grade 2 = red line | Grade 3 = purple line



dis_t	0.72	0.72	1	1									
miss_t	0.62	0.57	0.59	0.60	1	-							
add_t	0.54	0.55	0.58	0.57	0.66	0.62	1	1					
sub_t	0.51	0.49	0.49	0.52	0.75	0.72	0.68	0.71	1	1			
word_t*	0.59	0.43	0.55	0.48	0.65	0.68	0.57	0.58	0.63	0.64	1	1	

Note: All correlations significant at the  $p < 0.000$  level

Figure 5. *Grade 2 numeracy subtask distributions*

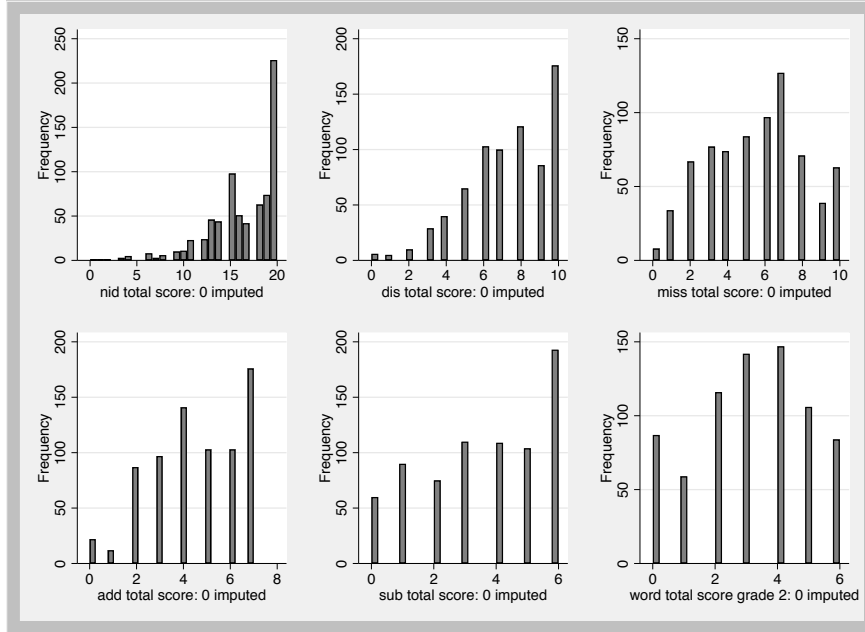
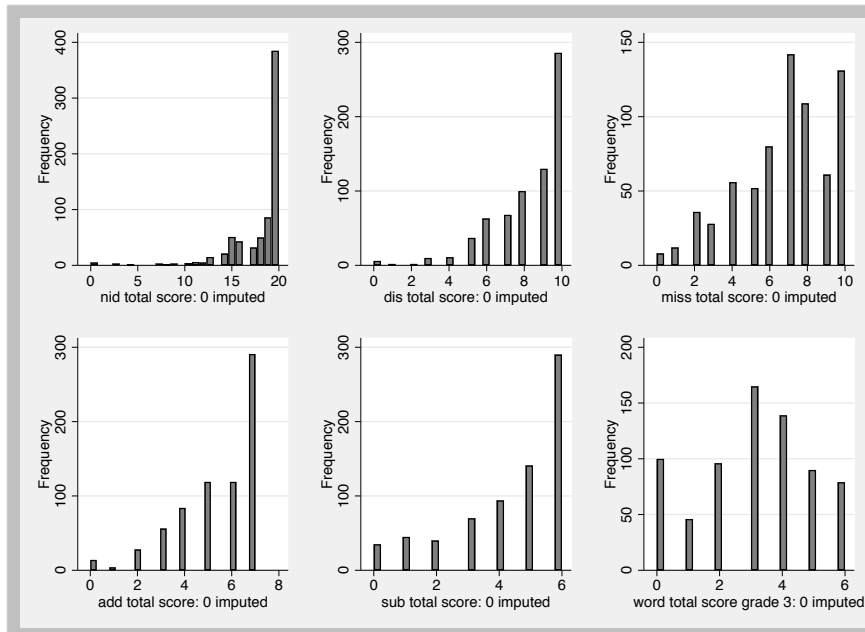


Figure 6. *Grade 3 numeracy subtask distributions*



Students in both grades 2 and 3 on average scored highly on the number identification (mean percent correct ( $MPC_{g2} = 82\%$ ,  $MPC_{g3} = 91\%$ ) and number discrimination subtasks ( $MPC_{g2} = 73\%$ ,  $MPC_{g3} = 83\%$ ), both of which displayed ceiling effects. Students in both grades 2 and 3 appeared to have more difficulty with the missing number ( $MPC_{g2} = 55\%$ ,  $MPC_{g3} = 68\%$ ) and word problem ( $MPC_{g2} = 52\%$ ,  $MPC_{g3} = 51\%$ ); the distribution of both subtasks' scores approximated a normal distribution, particularly in grade 2. Finally, children in grade 3 on average scored highly on addition ( $MPC_{g3} = 78\%$ ) and subtraction ( $MPC_{g3} = 73\%$ ) subtasks, with distributions showing evidence of ceiling effects. While children in grade 2 on average answered 66% and 60% of addition and subtraction items correctly, respectively, the distributions were negatively skewed. Children in grade 3 scored significantly higher than children in grade 2 on all common subtasks (number identification:  $t(1453) = 9.35$ ,  $p = 0.000$ ; number discrimination:  $t(1454) = 8.56$ ,  $p = 0.000$ ; missing number:  $t(1454) = 9.10$ ,  $p = 0.000$ ; addition:  $t(1454) = 9.12$ ,  $p = 0.000$ ; and subtraction:  $t(1454) = 7.89$ ,  $p = 0.000$ ).

Correlations between subtasks seem to cluster into two groups: Number identification and number discrimination scores are highly correlated while missing number, addition, subtraction, and word problems are highly correlated, with moderate correlations between subtasks in the two groups. This pattern of correlations differ in some ways from prior research with the EGMA in the Syrian response region(see see Appendix A, available upon request).

**Exploratory analyses.** Exploratory factor analyses using all numeracy subtasks were conducted separately with the grade 2 and the grade 3 exploratory samples. As shown in Table 11, the one-factor solution (Model A) did not provide a good fit to the data in either grade. The two-factor model (Model B) provided a better fit to the data in both grades, but only provided an acceptable fit in grade 3. Visual inspection of the factor loadings in grade 3 indicated that number identification and number discrimination loaded on to one factor while the remaining four subtasks loaded on to a second factor. Because a two-item factor presents challenges in interpretation, and because all of the items are moderately to highly intercorrelated, I then examined variations on the one-factor model. As the two highest residual correlations in both grades involved number discrimination, I tried removing number discrimination from the model. These models did not provide a good fit to the data. However, models in which I retained number discrimination and modeled the residual correlations provided an acceptable fit to the data in grade 2 and a good fit to the data in grade 3 (Model C; in grade 2, between number discrimination, number identification, and subtraction; in grade 3, between number identification and number discrimination, and between addition and

subtraction). Notably, in Model C in both grades, the residual correlation between number discrimination and subtraction was negative. Finally, I fit a one-factor model (Model D) to the correlations between missing number, addition, subtraction, and word problem subtasks. This model provided an acceptable fit to the data in grade 2 and a good fit to the data in grade 3, when allowing for a residual correlation between addition and subtraction. I concluded that a one-factor “numerical operations” model is acceptable for describing correlations among missing number, addition, subtraction, and word problem subtasks. However, based on the residual correlation between number identification and number discrimination familiar – as well as the improved model fit when number identification and discrimination are removed from the model – it is likely that there are two distinct but related early numeracy constructs in grades 2 and 3, as found in prior research.

Table 11. *Exploratory factor analysis model fit*

		Model A	Model B	Model C	Model D
Grade 2	X <sup>2</sup>	81.548*	17.775*	14.916	6.037
	df	9	4	6	2
	RMSEA	0.148	0.096	0.063	0.074 (0.005-
	(90% CI)	(0.119- 0.178)	(0.054- 0.144)	(0.023 0.105)	0.145)
	CFI	0.944	0.989	0.993	0.996
Grade 3	X <sup>2</sup>	87.73	11.680*	11.752	1.82
	df	9	4	7	1
	RMSEA	0.157	0.073 (0.026 -	0.044	0.048
	(90% CI)	(0.128-0.187)	0.124)	(0.000-0.086)	(0.000 -0.158)
	CFI	0.939	0.994	0.996	.999

Note: X<sup>2</sup> and df refer to the chi-square test off model fit and its degrees of freedom; RSMEA (90% CI) denotes the root mean square error of approximation and its 90% confidence interval; CFI denotes the comparative fit index.

**Confirmatory factor analyses.** While Model C provided a better fit to the data in grade 2, the three residual correlations – one of which was negative – would require the use of refined scoring methods not optimal for a formative assessment. I thus proceeded with confirming Model D using the grade 2 and 3 confirmatory sample. (I note that Model D also included a residual correlation between addition and subtraction in grade 3; however, it was a positive correlation likely attributable to addition and subtraction being administered within the same test subsection.) This model provided a good fit to the data in both grades 2 (X<sup>2</sup> = 1.04; RMSEA = 0.00 (0.00-.10); CFI = 1.00) and 3 (X<sup>2</sup> = .381; RMSEA = 0.00 (0.00-.12); CFI = 1.00). Table 12 shows the standardized factor loadings for each final confirmatory model. All items load highly on the underlying factor ( $\lambda > .700$ ). However, we emphasize that comparisons should not be drawn



between grades about the magnitude of factor loadings given that each model contains different subtasks due to both planned administration and data collection challenges.

Table 12. *Standardized actor loadings for confirmatory models*

	Grade 2	Grade 3
miss_t	.856	.864
add_t	.808	.761
sub_t	.905	.829
word_t	.761	.850
Residual correlation: addition-subtraction		.326

Note: All factor loadings were significant at  $p < .001$ , after adjusting for clustered standard errors.

\*Different versions administered in grades 2 and 3

## Reliability.

**Internal consistency.** Cronbach’s alpha coefficients of numeracy scores derived from the final confirmatory model ( $\alpha_{g2} = .85$ ,  $\alpha_{g3} = .88$ ) were above commonly accepted benchmarks for good internal consistency ( $\alpha > .80$ ).

**Interrater reliability.** Krippendorff’s alpha coefficients for each literacy subtask are shown in Table 8. All coefficients are above the common benchmark for substantial interrater agreement ( $\alpha > .80$ ).

Table 13. *Krippendorff’s alpha coefficients*

nid_t	dis_t	miss_t	add_t	sub_t	word_t*
.89	.96	.99	.98	.93	.97

**Item difficulty.** Fit statistics for unidimensional confirmatory factor models fit to item-level subtask data are reported in Appendix Table 1. For subtasks in which a one-factor model provided a good fit to the data, a comparison between 1- and 2-PL model fit is provided in Appendix Table 2. For subtasks with evidence of unidimensionality, Appendix Figures 1 – 4 report item difficulty parameters from 2-PL models fit to item-level number discrimination (grades 2 and 3); number discrimination (grades 2 and 3); addition (grade 3) and subtraction (grade 3) data.

We refer the interested reader to these figures for more details on individual item difficulty. Broadly, however, we note that the numeracy subtask data we examined using the IRT approach provided information about children with average to below average number identification, number discrimination, addition and subtraction skill

levels. In addition, visual inspection of the empirical item difficulties indicated general convergence with the hypothesized item difficulties (see “Measure” section for more detail).

## Social and Emotional

Given that the same versions of social and emotional subtasks were administered in each grade, results of analyses are reported for the combined grade 2 + grade 3 samples.

**Descriptive statistics.** Table 13 contains descriptive statistics for raw item scores while Table 14 displays the tetrachloric correlation matrix among item scores. Given the known inconsistencies in application of stop and skip rules, we report here item responses with missing data.

Table 13. *Social and emotional item descriptive statistics*

Variable	Obs	Mean	Std. Dev.	Min	Max
Empathy - sad					
emps_m1	1,442	97.80%	0.15	0	100%
emps_m2	1,325	95.32%	0.21	0	100%
emps_m3	926	73.97%	0.44	0	100%
emps_m4	1,331	77.46%	0.42	0	100%
emps_m5	1,333	84.25%	0.36	0	100%
emps_23	1,327	1.46	0.58	0	2
Empathy - anger					
empa_m1	1,429	93.70%	0.24	0	100%
empa_m2	1,269	90.15%	0.30	0	100%
empa_m3	966	68.12%	0.47	0	100%
empa_m4	1,314	74.28%	0.44	0	100%
empa_m5	1,324	78.40%	0.41	0	100%
empa_23	1,278	1.41	0.66	0	2
Empathy - worry					
empw_m1	1,213	42.13%	0.49	0	100%
empw_m2	1,012	70.26%	0.46	0	100%
empw_m3	759	50.33%	0.50	0	100%
empw_m4	1,134	55.56%	0.50	0	100%
empw_m5	1,125	59.38%	0.49	0	100%
empw_23	1,024	1.07	0.80	0	2
Perseverance					
per_m1	1,439	90.41%	0.29	0	100%
per_m2	1,422	87.48%	0.33	0	100%
per_m3	1,384	76.52%	0.42	0	100%
per_m4	1,350	67.78%	0.47	0	100%

# Self-concept

self_m2	1,405	97.01%	0.17	0	100%
self_m3	1,197	71.93%	0.45	0	100%
self_m4	1,341	92.10%	0.27	0	100%
self_m5	1,005	76.22%	0.43	0	100%
self_m6	897	59.87%	0.49	0	100%
self_m7	969	72.34%	0.45	0	100%
self_25	1,406	1.51	0.54	0	2
self_36	1,225	1.14	0.81	0	2
self_47	1,343	1.44	0.62	0	2

Figures 7 and 8. *Empathy, perseverance, and self-concept item distributions*

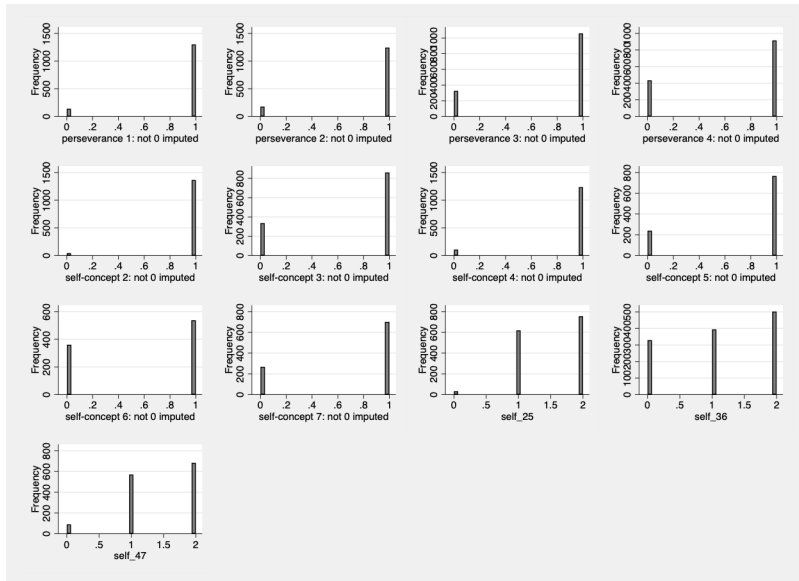
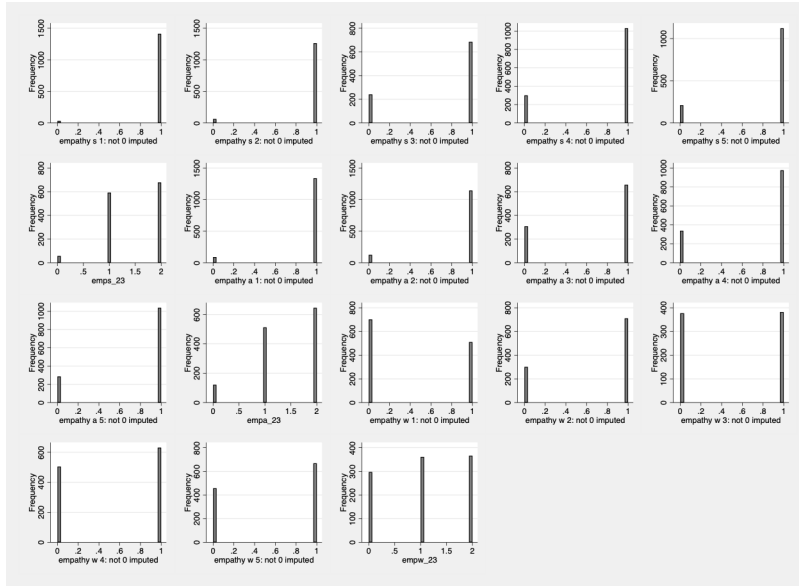


Table 14. *Social and emotional item tetrachloric correlation matrix*

	EMPS_ _M1	EMPS_ _M4	EMPS_ _M5	EMPA_ _M1	EMPA_ _M4	EMPA_ _M5	EMPW_ _M1	EMPW_ _M4	EMPW_ _M5	PER_ M1	PER_ M2	PER_ M3	PER_ M4	EMPS_ _23	EMPA_ _23	EMPW_ _23	SELF_ _25	SELF_ _36
EMPS_ M1																		
EMPS_ M4	0.525																	
EMPS_ M5	0.546	0.817																
EMPA_ M1	<b>0.387</b>	0.128	0.185															
EMPA_ M4	0.093	<b>0.598</b>	0.518	0.604														
EMPA_ M5	0.093	0.433	<b>0.623</b>	0.626	0.876													
EMPW_ _M1	0.11	0.095	0.001	<b>0.268</b>	0.297	0.198												
EMPW_ _M4	0.265	<b>0.519</b>	0.498	0.368	<b>0.728</b>	0.693	0.491											
EMPW_ _M5	0.304	0.409	<b>0.551</b>	0.38	0.631	<b>0.734</b>	0.435	0.946										
PER_M 1	0.147	0.239	0.269	0.086	0.127	0.223	-0.259	0.015	0.002									
PER_M 2	0.233	0.313	0.379	0.162	0.15	0.327	-0.207	-0.007	0.053	0.897								
PER_M 3	0.088	0.246	0.419	0.295	0.258	0.448	-0.075	0.16	0.228	0.669	0.835							
PER_M 4	0.028	0.199	0.438	0.125	0.224	0.486	0.02	0.207	0.338	0.708	0.781	0.919						
EMPS_ 23	0.588	0.414	0.453	0.284	0.228	0.387	0.171	0.281	0.258	0.231	0.334	0.303	0.299					
EMPA_ 23	0.328	0.334	0.49	0.702	0.581	0.698	0.289	0.5	0.577	0.106	0.282	0.444	0.487	<b>0.595</b>				
EMPW_ _23	0.284	0.236	0.262	0.378	0.453	0.524	0.752	0.745	0.801	-0.189	-0.068	0.23	0.393	<b>0.447</b>	<b>0.681</b>			
SELF_2 5	0.042	0.365	0.256	0.151	0.456	0.426	0.215	0.287	0.234	0.377	0.381	0.388	0.395	0.306	0.403	0.266		
SELF_3 6	-0.23	0.285	0.272	0.042	0.326	0.337	0.289	0.384	0.389	0.235	0.255	0.304	0.341	0.354	0.427	0.395	0.765	
SELF_4 7	-0.023	0.343	0.38	0.219	0.473	0.502	0.245	0.431	0.455	0.296	0.358	0.457	0.573	0.398	0.573	0.474	0.843	0.761

Note: Items in bold are significant at the  $p < 0.05$  level.



***Empathy – emotion identification.*** Nearly all children were correctly able to the emotion in a drawing of a sad child and over 90 percent of children were able to correctly identify the emotion in a picture of an angry child. However, less than half of children could correctly identify the emotion in a photo of a child with a worried expression. Item responses were modestly correlated ( $0.11 < r < 0.39$ ).

***Empathy – pro-social response.*** The majority of children were able to identify at least one response to make a sad or angry child feel better, but only 70 percent of children who correctly identified that a child was worried was able to identify a response to make the child feel better. The three Guttman scale pro-social response items were moderately correlated ( $0.45 < r < 0.68$ ).

***Empathy hostile attribution items.*** Approximately 1/4 of children who listened to a vignette about an ambiguous social situation attributed hostility to the action of the child that caused the sad/angry response from the pictured child. Nearly half of children who listened to a vignette about an ambiguous social situation attributed hostility to the action of the child that cause the worried response from the pictured child. The tendency to attribute hostile intent in the sad scenario was moderately correlated with the tendency to attribute hostile intent in the anger and worry scenarios; the tendency to attribute hostile intent in the anger scenario was highly correlated with the response in the worry scenario.

***Empathy perspective taking:*** A majority of children were able to identify an appropriate emotion response of the child responsible for making the pictured child feel sad or angry, but only half of children could identify an appropriate emotion felt by the child responsible for making the pictured child feel worried. The three perspective taking items were moderately to highly correlated ( $0.55 < r < 0.73$ ).

***Perseverance.*** 64 percent of children who attempted any of the perseverance drawing tasks completed all 4 tasks. 5.5% of children who attempted the task did not complete one drawing and 11.27% of children did not complete two drawings. The four items within the perseverance subtask were highly correlated ( $r > .65$ ).

***Self-concept.*** A majority of children were able to identify at least one thing s/he hoped to happen in the future. On average children were able to identify one thing that could prevent them from that hoped-for-self and more than one thing that could

support them in achieving their goal. The three Guttman scale items within self-concept were highly correlated ( $r > .75$ )

**Exploratory factor analysis.** Exploratory factor analyses using all social and emotional items were conducted with the combined grade 2 and grade 3 exploratory samples. Missing item-level data was handled using a WLSMV with pairwise present estimator in MPlus 8.0, which – similar to full-information maximum likelihood approaches -- estimates correlations using all available data. As shown in Table 15 below, the correlations between items are best modelled by four latent constructs: a self-concept construct (Factor D), a perseverance construct (Factor B), a factor that contains many of the empathy sadness items (Factor A) and a factor that contains a mix of empathy anger and worry items (Factor B). Interestingly, factors are only modestly correlated, although factor correlations are likely to increase with solutions in items are not allowed to cross-load.

We next removed the perseverance and self-concept items from the model in order to further examine the factor structure of the empathy items. We explored a number of solutions: (i) extracting one to six factors with all items; (ii) removing the emotion identification worry item and extracting one to six factors; (iii) removing all empathy worry items and extracting one to four factors; (iv) testing models in which we allowed for “methods” factors that account for variance due to items being nested within emotions. A majority of these models did not provide a good fit to the data; of those that did, the pattern of factor loadings did not make conceptual sense. All results of this exploratory analysis are available upon request; we provide our interpretation of these findings and next steps in the discussion section.

**Confirmatory factor analysis.** I proceeded with confirming a two-factor model in which perseverance items loaded onto one factor and self-concept items loaded onto a second factor. This model provided a good fit to the data ( $\chi^2 = 1.89$ ; RMSEA = 0.00 (0.00-.10); CFI = 1.00). Table 16 shows the standardized factor loadings for each final confirmatory model. All items load highly on the underlying factor ( $\lambda > .700$ ), and factors were moderately correlated.

Table 15. *Exploratory factor analysis models*

		Factor A	Factor B	Factor A	Factor B	Factor C	Factor A	Factor B	Factor C	Factor D
Factor										
loadings	EMPS_M1	0.421*	-0.042	0.569*	0.092	-0.287	0.520*	0.09	0.152	-0.27
	EMPS_M4	0.640*	0.196*	0.771*	0.266*	-0.230*	0.900*	0.005	-0.205	0.152*
	EMPS_M5	0.698*	0.216*	0.870*	0.333*	-0.288*	0.895*	0.150*	-0.042	-0.002
	EMPA_M1	0.644*	0.004	0.560*	-0.059	0.190*	0.047	0.228*	0.738*	-0.158
	EMPA_M4	0.768*	0.135*	0.768*	0.082	0.098	0.567*	0.097	0.380*	0.066
	EMPA_M5	0.756*	0.291*	0.781*	0.261*	0.122*	0.522*	0.342*	0.480*	-0.007
	EMPW_M1	0.557*	-0.235*	0.277*	-0.454*	0.460*	-0.157	-0.262*	0.645*	0.219*
	EMPW_M4	1.000*	-0.175*	0.936*	-0.279*	0.115	0.640*	-0.296*	0.505*	0.161*
	EMPW_M5	0.965*	-0.100*	0.902*	-0.193*	0.135*	0.550*	-0.141*	0.575*	0.095
	PER_M1	-0.255*	0.892*	-0.044	0.887*	-0.015	0.128	0.798*	-0.308*	0.091
	PER_M2	-0.118*	0.926*	0.091	0.923*	-0.009	0.193*	0.885*	-0.182*	0.033
	PER_M3	-0.014	0.953*	0.066	0.868*	0.248*	-0.045	0.936*	0.174*	0.042
	PER_M4	0.01	0.929*	0.041	0.826*	0.344*	-0.130*	0.899*	0.224*	0.156*
	EMPS_23	0.335*	0.273*	0.316*	0.227*	0.178*	0.190*	0.247*	0.189*	0.155*
	EMPA_23	0.750*	0.220*	0.624*	0.095	0.365*	0.142	0.326*	0.707*	0.08
	EMPW_23	0.822*	-0.078	0.549*	-0.300*	0.528*	-0.016	-0.069	0.837*	0.216*
	SELF_25	0.237*	0.782*	-0.039	0.213*	0.862*	0.047	0.049	-0.126*	0.944*
	SELF_36	0.278*	0.660*	0.027	0.149*	0.802*	0.014	0.014	0.003	0.864*
	SELF_47	0.346*	0.768*	0.088*	0.182*	0.870*	-0.024	0.119*	0.156*	0.863*
Factor										
correlations	Factor B	0.261*		0.150*				0.221*		
	Factor C			0.343*	0.150*			0.331*	0.039	
	Factor D							0.304*	0.285*	0.303*
Fit	X <sup>2</sup>	762.529*		504.515*				328.397*		
	df	134		117				101		
	RMSEA	0.08		0.068				0.056		
	RMSEA CI	0.075 0.086		0.062 0.074				0.049 0.062		
	CFI	0.915		0.948				0.969		
	TLI	0.892		0.924				0.948		



Table 16. Standardized actor loadings for confirmatory models

	Perseverance	Self-concept
per_1	.897	
per_2	.941	
per_3	.927	
per_4	.946	
self_1		.864
self_2		.720
self_3		.957
per-self		.532

Note: All factor loadings were significant at  $p < .001$ , after adjusting for clustered standard errors.

### Reliability.

**Internal consistency.** I examined the internal consistency of the hypothesized subscales by calculating Cronbach's alpha. While the emotion identification scale alpha ( $\alpha = 0.167$ ) was very low, likely due to the difficulty with the worry item, perseverance ( $\alpha = 0.780$ ) and self-concept ( $\alpha = 0.80$ ) alphas were both high. In addition, for three item scales that likely contained measurement error due to inconsistencies in coding stop/skip patterns, the pro-social response ( $\alpha = 0.661$ ), hostile attribution bias ( $\alpha = 0.650$ ), and perspective-taking alphas ( $\alpha = 0.659$ ) were better than anticipated.

**Interrater reliability.** Krippendorff's alpha coefficients for each social and emotional items are shown in Table 17. All coefficients are above the common benchmark for substantial interrater agreement ( $\alpha > .80$ ).

Table 17. Krippendorff's alpha coefficients

emps_1	emps_23	emps_4	emps_5	per_	per_2	per_3	per_4	self_1	self_2	self_3
.89	.94	.96	.98	.74	.93	.94	1.00	.82	.92	.93

## Discussion

In this report, we sought to provide evidence on whether data from the HAL instrument could be interpreted as providing consistent and meaningful information about children's literacy, numeracy, and social and emotional skills for formative feedback purposes in the Whole of Syria response. Overall we conclude that there is evidence to support the interpretation of scores as assessing early grade literacy, numeracy, and some social and emotional skills; that scores within each domain are internally consistent; and that data collectors can provide consistent scores about children's literacy, numeracy, and social-emotional skills. There is also evidence that data from this assessment may be better fit to purpose than that collected using prior international early grade learning assessments. We summarize the key findings and convergences and divergences with evidence from prior assessment efforts below.

### **Moving from what children do not know to what children *do* know**

Compared to analyses undertaken with data collected previously using the EGRA and EGMA in Lebanon and Syria (see Appendix A, available upon request), students on average correctly responded to a greater percentage of items on common literacy and numeracy subtasks in the HAL. Fewer floor effects were observed in HAL literacy and numeracy subtasks, and indeed, many HAL subtasks had the opposite issue: ceiling effects. Item difficulty analyses conducted with letter sound, familiar word, number identification, and number discrimination items further support our conclusion that the instrument is providing information about what children do know; they also demonstrate that it is possible to develop early grade literacy and numeracy items that provide information about children whose underlying skill level is at least 2 standard deviations below the mean. In the context of conflict and crisis, where interruptions in schooling can impede children's development of fundamental academic and social-emotional skills, it is critical to have access to assessments that can provide such information, particularly for formative purposes: Information about the skills children *do* have can guide teachers in taking a strengths-based approach to scaffolding further learning.

### **The importance of contextually appropriate assessments**

There are, of course, several explanations for the difference in response distributions between samples. It could be, for example, that students in the HAL sample had relatively fewer disruptions in their education due to conflict and migration than students assessed in Syria and Lebanon two years prior. However, it is also possible that

items on this assessment are better able to capture students' academic skills due to carefully taking into account during the process of test construction research on Arabic literacy development as well as the Syrian curriculum. For example, items in the letter sound and familiar word subtask were ordered based on hypothesized match in sound between MSA and Syrian dialect, frequency of occurrence in Syrian dialect, and written presentation. Visual inspection of the empirical item difficulties provided some support for the hypothesized difficulties – particularly for letter sound item responses in both grades 2 and 3 and familiar word reading items in grade 2 – which poses some intriguing possibilities for early literacy assessment development and administration. For instance, the letter sound task administered as part of the EGRA in Syria does not contain diacritics, vowel markers which provide critical guidance as students are learning to read MSA. In the HAL, items varied in whether they contain diacritics or not: Items four and item five are the same letter, but item four contains a diacritic and item five does not. Item four is the fourth easiest item for children in grades 2 and 3 to answer while item five is the *forty-second* easiest item for children to answer – out of 50. In addition, while items in the EGRA are not directly comparable to the HAL given both the lack of diacritics and the different letter forms included in the EGRA, it is notable that eight out of 10 of the letters included in the first row of Syrian EGRA – the stopping rule cutoff -- are answerable by children with an underlying skill level at -.5 to -1 SD below the mean. In contrast, half of the letter sound items on the HAL are answerable by children whose skill level is -1 SD below the mean or lower.

While we suggest that these promising literacy and numeracy results empirically underscore the need for more contextually appropriate assessments – and the processes for constructing them -- we also recognize the consequences of not paying enough attention to adaptation in the test development prices. For example, as part of the empathy subtask children were asked to identify the emotions of Caucasian children in three pictures. Research indicates that the match between the cultural background of the individual expressing an emotion in a still photograph and the judge is important for emotion recognition accuracy – and particularly important for more complex emotions such as worry (see, e.g., meta-analysis by Elfenbein and Ambady, 2002). This suggests that it is likely problematic to interpret scores on the emotion identification items within the empathy subtask as providing consistent and meaningful information. Moreover, given that the pictures were the first question within a five-item “testlet” and that stop rules directed enumerators not to administer the subsequent four items if the child could not correctly identify the emotion, it is likely that the following scores (pro-social response, hostile attribution bias, perspective taking) contain error which could

preclude our ability to identify a factor model that provided a good fit to the data. Future versions of the HAL will address these limitations.

### **An opportunity for learning**

Finally, we emphasize that through the process of developing the HAL assessment we had a critical opportunity to pose and test hypotheses about literacy, numeracy, and social and emotional development in an Arab context. In turn, such inquiry can open new approaches to both education programming and assessments in such contexts. In terms of programming, for example, it is striking in the HAL results that children in both grades 2 and 3 who reported knowing that in Arabic there was a *fusha* and *ammiya* scored significantly higher on *all* literacy subtasks. As highlighted in the work of colleague contributors, it may be particularly important for early grade teachers to explicitly explain and call attention to the diglossic feature of the Arabic language to help children orient and make sense of the differences in what they are hearing, speaking, and reading (Dakwar, 2005; Khamis-Dakwar & Khattab, 2014). For those developing assessments for use in such contexts, being able to capture information about what children do know as well as what children can ultimately help develop guidance for scoring that provides more meaningful information. For example, based on the residual correlation between familiar word reading and letter sound identification – as well as the improved model fit when letter sound is removed from the model – it is possible that there may be two distinct but related early literacy constructs in Arabic (perhaps a decoding skill and a comprehension skill). This is different than the one-factor structure of early Arabic literacy skill suggested by our secondary analysis work, in which the large number of zero scores likely inflated the correlation between subtasks. If so, this would inform the design of future assessments and suggest different avenues for how teachers should score results.

### **Limitations**

We note two limitations in addition to those discussed above. First, there were challenges in the consistent application of stop and skip rules – as well as in data processing, cleaning, and verification – that likely added additional error into the data (further details available on request). Second, as described in the analytic plan section, our subtask-level factor analyses (RA 2 and 3) of literacy and numeracy data makes the assumption that all items within a subtask discriminate to the same degree between underlying ability levels (within IRT framework) or load equally onto a latent subtask construct (within a factor analytic framework). As shown by our IRT analyses, this assumption does not hold for all subtasks; in addition, follow-up examination of empirical item fit indicated that half of the letter sound items, for example, did not

provide a good fit to the 1-PL model. However, meeting the assumptions required for unweighted sum scoring of item-level data would require removing those items, which then limits the information available to the teacher to use for feedback purposes. This may prove particularly restrictive in Arabic, given variations in letter form, vowelization, and orthography. Given that the psychometric criteria required for formative assessments are lower than those required for assessments used for higher-stakes purposes – as well as the challenges to providing student feedback lacking a breadth of item-level information – we recommend retaining items and proceeding with unweighted sum scoring.

## **Conclusion**

The decision on how to approach item-level scoring of subtasks was driven by triangulation of psychometric concerns with education practice concerns and child developmental considerations. We believe that this process of triangulation holds great promise for the development of field feasible, contextually relevant, and rigorous assessments, as shown by the results of the pilot HAL assessment tool.

## Recommendations Based on Psychometric Analyses

### For revisions to test content for formative purposes

**Increase the difficulty level of subtasks with ceiling effects.** This includes letter sound identification, familiar word reading, listening comprehension, vocabulary, number discrimination, and number identification in all grades.

**Create different version of addition and subtraction subtasks for grades 2 and 3.** While the distribution of addition and subtraction scores was negatively skewed in both Grades 2 and 3, the distributions showed ceiling effects in Grade 3. This suggests that given the difference in average skill level between grades, two different versions of the subtasks should be created.

**Replace pictures in empathy subtask with pictures matching the cultural background of children.** Research suggests that the match in cultural background of an individual expressing an emotion in a still photograph and the judge is important for emotion recognition accuracy (see, e.g., meta-analysis by Elfenbein and Ambady, 2002). This error in the scores could in part account for difficulties fitting a factor analysis model to the empathy subtask scores.

**Consider revising the emotions children are asked to identify.** Children had difficulty identifying and responding to the worry picture and scenario, consistent with prior research suggesting that worry is one of the harder emotions to recognize when presented with a picture of an outgroup member.

### For revisions to test administration for formative purposes

**Ensure clarity in stop and skip rule instructions and consistency in application.** In particular, specify directions as to whether non-attempted items should be recoded to incorrect or missing when a child stops each subtask.

**Capture information about attempted items.** We recommend recording correct and incorrect attempts at items in the oral passage reading subtask (where scores of those unable to correctly respond to five items in 30 seconds are currently recoded to zero, resulting in a loss of information).

### For scoring for formative purposes

**Literacy.** We recommend scoring familiar word reading, oral passage reading, reading comprehension, and vocabulary subtask together, given that correlations between most subtasks are consistent with a single, latent early literacy construct. We also recommend scoring listening comprehension and letter sound identification subtasks separately.

**Numeracy.** We recommend scoring number identification and number discrimination together and missing number, addition, subtraction, and word problems together, given that the factor analysis results provided evidence to suggest two related but distinct early numeracy constructs. We also recommend recording answers to both questions asked in the new word problem format. Right now, children who are able to comprehend/plan but not arrive at the correct answers are coded the same as children who cannot comprehend.

**Social and emotional.** We recommend scoring perseverance items together and self-concept items together.

## References

- 3EA. (2018). *3EA TIES-IRC Measurement Tools: Children's Holistic Learning and Development (CHILD) and Program Implementation Quality*. New York, NY: NYU Global TIES for Children.
- Aber, J. L., Tubbs, C., Torrente, C., Halpin, P. F., Johnston, B., Starkey, L., ... Wolf, S. (2017). Promoting children's learning and development in conflict-affected countries: Testing change process in the Democratic Republic of the Congo. *Development and Psychopathology*, 29(1), 53–67.  
<https://doi.org/10.1017/S0954579416001139>
- Al-Krenawi, A., & Graham, J. R. (2012). The impact of political violence on psychosocial functioning of individuals and families: the case of palestinian adolescents. *Child and Adolescent Mental Health*, 17(1), 14–22.  
<https://doi.org/10.1111/j.1475-3588.2011.00600.x>
- Bartlett, L., Dowd, A. J., & Jonason, C. (2015). Problematizing early grade reading: Should the post-2015 agenda treasure what is measured? *International Journal of Educational Development*, 40, 308–314.  
<https://doi.org/10.1016/j.ijedudev.2014.10.002>
- 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](https://doi.org/10.1207/s15328007sem1302_2)
- Bock, R. D., & Aitken, M. (1981). Marginal maximum likelihood estimation of item parameters: Application of an EM algorithm. *Psychometrika*, 46.
- Caires, R., & Tubbs Dolan, C. (2018). *Supporting decision-making for children's programming in crisis contexts*. Retrieved from NYU Global TIES for Children website:  
[https://steinhardt.nyu.edu/scmsAdmin/media/users/mhm327/Survey\\_Memo-4\\_FNL\\_updated.pdf](https://steinhardt.nyu.edu/scmsAdmin/media/users/mhm327/Survey_Memo-4_FNL_updated.pdf)
- Cronbach, L. J., & Shavelson, R. J. (2004). My Current Thoughts on Coefficient Alpha and Successor Procedures. *Educational and Psychological Measurement*, 64(3), 391–418. <https://doi.org/10.1177/0013164404266386>
- Cummings, E. M., Merrilees, C. E., Taylor, L. K., & Mondy, C. F. (2017). Developmental and social–ecological perspectives on children, political violence, and armed conflict. *Development and Psychopathology*, 29(1), 1–10.  
<https://doi.org/10.1017/S0954579416001061>
- Dakwar, R. K. (2005). Children's attitudes towards the diglossic situation in Arabic and its impact on learning. *Languages, Communities, and Education*, 3, 75–86.



- DiStefano, C., & Zhu, M. (2009). Understanding and using factor scores: Considerations for the applied researcher. *Practical Assessment, Research & Evaluation*, 14(20), 1–11.
- Dodge, K. A., Malone, P. S., Lansford, J. E., Sorbring, E., Skinner, A. T., Tapanya, S., ... others. (2015). Hostile attributional bias and aggressive behavior in global context. *Proceedings of the National Academy of Sciences*, 112(30), 9310–9315.
- Dowd, A. J., & Bartlett, L. (2019). The Need for Speed: Interrogating the Dominance of Oral Reading Fluency in International Reading Efforts. *Comparative Education Review*, 63(2), 189–212. <https://doi.org/10.1086/702612>
- Dryden-Peterson, S. (2009). *Barriers to accessing primary education in conflict-affected fragile states*. Retrieved from Save the Children website: [http://www.toolkit.ineesite.org/resources/ineecms/uploads/1150/R2\\_Dryden-Peterson.pdf](http://www.toolkit.ineesite.org/resources/ineecms/uploads/1150/R2_Dryden-Peterson.pdf)
- Dryden-Peterson, Sarah. (2011). *Refugee education: A global review*. Retrieved from UNHCR website: <http://www.unhcr.org/en-us/4fe317589.pdf>
- Eisenberg, N., Zhou, Q., & Koller, S. (2001). Brazilian Adolescents' Prosocial Moral Judgment and Behavior: Relations to Sympathy, Perspective Taking, Gender-Role Orientation, and Demographic Characteristics. *Child Development*, 72(2), 518–534. <https://doi.org/10.1111/1467-8624.00294>
- Embretson, S. E., Reise, S. P., & Reise, S. P. (2013). *Item Response Theory*. <https://doi.org/10.4324/9781410605269>
- Ferguson, C. A. (1959). Diglossia. *Word*, 15(2), 325–340.
- Ford, C. B., Kim, H. Y., Brown, L., Aber, J. L., & Sheridan, M. A. (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), 141–157. <https://doi.org/10.1177/1745499919829217>
- Gabrieli, C., Ansel, D., & Krachman, S. (2015). *Ready to be counted: The research case for education policy action on non-cognitive skills*. Boston, MA: Transforming Education.
- Goodman, R. (1997). The Strengths and Difficulties Questionnaire: A Research Note. *Journal of Child Psychology and Psychiatry*, 38(5), 581–586. <https://doi.org/10.1111/j.1469-7610.1997.tb01545.x>
- Hadi, F. A., & Llabre, M. M. (1998). The Gulf crisis experience of Kuwaiti children: Psychological and cognitive factors. *Journal of Traumatic Stress*, 11(1), 45–56. <https://doi.org/10.1023/A:1024453015176>
- 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>
- 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.
- Hambleton, R. K., Swaminathan, H., & Rogers, H. J. (1991). *Fundamentals of item response theory* (Vol. 2). Sage.
- Hayes, A. F., & Krippendorff, K. (2007). Answering the Call for a Standard Reliability Measure for Coding Data. *Communication Methods and Measures*, 1(1), 77–89. <https://doi.org/10.1080/19312450709336664>
- 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>
- INEE. (2016). *Psychosocial support and social and emotional learning for children and youth in emergency settings* [Background paper]. Retrieved from INEE website: [http://toolkit.ineesite.org/resources/ineecms/uploads/1126/20161219\\_PSS\\_SEL\\_Background\\_Note\\_Digital\\_Final.pdf](http://toolkit.ineesite.org/resources/ineecms/uploads/1126/20161219_PSS_SEL_Background_Note_Digital_Final.pdf)
- Inter-Agency Standing Committee. (2007). *IASC Guidelines for mental health and psychosocial support in emergency settings*. Retrieved from IASC website: [http://www.who.int/mental\\_health/emergencies/9781424334445/en/](http://www.who.int/mental_health/emergencies/9781424334445/en/)
- International Association for the Evaluation of Educational Achievement. (n.d.). Trends in International Mathematics and Science Study (TIMSS) - TIMSS Participating Countries. Retrieved June 15, 2019, from <https://nces.ed.gov/timss/countries.asp>
- International Rescue Committee. (2017). *Impact of war on Syrian children's learning*. Retrieved from Author website: <https://www.rescue.org/sites/default/files/document/1434/educationreportlearninglevelssyrianchildrenfinal.pdf>
- Jones, D. E., Greenberg, M., & Crowley, M. (2015). Early Social-Emotional Functioning and Public Health: The Relationship Between Kindergarten Social Competence and Future Wellness. *American Journal of Public Health*, 105(11), 2283–2290. <https://doi.org/10.2105/AJPH.2015.302630>
- Jordans, M. J. D., Pigott, H., & Tol, W. A. (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), 9. <https://doi.org/10.1007/s11920-015-0648-z>
- Jordans, M. J., Tol, W. A., Susanty, D., Ntamatumba, P., Luitel, N. P., Komproe, I. H., & de Jong, J. T. (2013). Implementation of a mental health care package for

- children in areas of armed conflict: a case study from Burundi, Indonesia, Nepal, Sri Lanka, and Sudan. *PLoS Med*, 10(1), e1001371.
- Khamis, V. (2019). Posttraumatic stress disorder and emotion dysregulation among Syrian refugee children and adolescents resettled in Lebanon and Jordan. *Child Abuse & Neglect*, 89, 29–39. <https://doi.org/10.1016/j.chiabu.2018.12.013>
- Khamis-Dakwar, R., & Froud, K. (2014). Neurocognitive modeling of the two language varieties in Arabic Diglossia. *Perspectives on Arabic Linguistics XXVI: Papers from the Annual Symposium on Arabic Linguistics. New York, 2012, 2*. John Benjamins Publishing Company.
- Khamis-Dakwar, R., & Khattab, G. (2014). Cultural and linguistic considerations in language assessment and intervention for Levantine Arabic speaking children. *Perspectives on Communication Disorders and Sciences in Culturally and Linguistically Diverse (CLD) Populations*, 21(3), 78–87.
- Kim, H. Y., Brown, L., Ferrans, S. D., & Weiss Yagoda, J. (2019). *The impact of IRC's Healing Classrooms tutoring on children's learning and social-emotional outcomes in Niger* (Policy Brief No. 2.2.2). Retrieved from International Rescue Committee website: [https://steinhardt.nyu.edu/scmsAdmin/media/users/nh1375/3EA-Niger\\_Policy\\_Brief\\_updated\\_3.2019\\_A.pdf](https://steinhardt.nyu.edu/scmsAdmin/media/users/nh1375/3EA-Niger_Policy_Brief_updated_3.2019_A.pdf)
- Kim, H. Y., Brown, L., Tubbs Dolan, C., Sheridan, M. A., & Aber, J. L. (2019). Post-migration risk factors, developmental processes, and learning outcomes among Syrian refugee children in Lebanon. *Journal of Applied Developmental Psychology*.
- Kline, R. B. (2011). *Principles and Practice of Structural Equation Modeling* (3rd Edition). New York, NY: Guilford.
- Krippendorff, K. (2011). *Computing Krippendorff's Alpha-Reliability* [Working paper]. Retrieved from University of Pennsylvania website: [https://repository-upenn-edu.proxy.library.nyu.edu/asc\\_papers/43/](https://repository-upenn-edu.proxy.library.nyu.edu/asc_papers/43/)
- Lei, P.-W. (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>
- Lord, F. M. (2012). *Applications of Item Response Theory To Practical Testing Problems*. <https://doi.org/10.4324/9780203056615>
- Marsh, H. W., Morin, A. J. S., Parker, P. D., & Kaur, G. (2014). Exploratory Structural Equation Modeling: An Integration of the Best Features of Exploratory and Confirmatory Factor Analysis. *Annual Review of Clinical Psychology*, 10(1), 85–110. <https://doi.org/10.1146/annurev-clinpsy-032813-153700>

- Masten, A. S., & Narayan, A. J. (2012). Child development in the context of disaster, war, and terrorism: pathways of risk and resilience. *Annual Review of Psychology*, 63, 227–257. <https://doi.org/10.1146/annurev-psych-120710-100356>
- Masten, A. S., Roisman, G. I., Long, J. D., Burt, K. B., Obradović, J., Riley, J. R., ... Tellegen, A. (2005). Developmental cascades: linking academic achievement and externalizing and internalizing symptoms over 20 years. *Developmental Psychology*, 41(5), 733.
- Montjourides, P. (2013). Education data in conflict-affected countries: The fifth failure? *PROSPECTS*, 43(1), 85–105. <https://doi.org/10.1007/s11125-012-9260-8>
- Montoya, S. (2018, July 18). Meet the SDG 4 data: Measuring how much children are learning. Retrieved June 15, 2019, from Global Partnership for Education website: <https://www.globalpartnership.org/blog/meet-sdg-4-data-measuring-how-much-children-are-learning>
- Muthén, B. O., & Muthén, L. K. (2014). MPlus (Version 7.2) [Computer software]. Los Angeles, CA: Muthén & Muthén.
- Muthén, L. K., & Muthén, B. O. (2012). *Mplus Version 7 user's guide*. Los Angeles, CA: Muthén & Muthén.
- No Lost Generation. (2019). *Investing in the future: Protection and learning for all Syrian children and youth*. Retrieved from <https://reliefweb.int/sites/reliefweb.int/files/resources/68406.pdf>
- Oburu, P. O., & Palmérus, K. (2016). Stress Related Factors among Primary and Part-Time Caregiving Grandmothers of Kenyan Grandchildren. *The International Journal of Aging and Human Development*. <https://doi.org/10.2190/XLQ2-UJEM-TAQR-4944>
- Osborn, J., & Fitzpatrick, D. (2012). Replication analysis in exploratory factor analysis: What it is and why it makes your analysis better. *Practical Assessment, Research & Evaluation*, 17(15), 1–8.
- Panter-Brick, C., Hadfield, K., Dajani, R., Eggerman, M., Ager, A., & Ungar, M. (2017). Resilience in Context: A Brief and Culturally Grounded Measure for Syrian Refugee and Jordanian Host-Community Adolescents. *Child Development*. <https://doi.org/10.1111/cdev.12868>
- PASEC. (2017, June 21). Evaluation internationale PASEC2019. Retrieved June 15, 2019, from Pasec website: <http://www.pasec.confemen.org/evaluation/evaluation-internationale-pasec2019/>
- Pat-Horenczyk, R., Peled, O., Miron, T., Brom, D., Villa, Y., & Chemtob, C. M. (2007). Risk-Taking Behaviors Among Israeli Adolescents Exposed to Recurrent Terrorism: Provoking Danger Under Continuous Threat? *American Journal of Psychiatry*, 164(1), 66–72. <https://doi.org/10.1176/ajp.2007.164.1.66>

- Perkins, J. D., Ajeeb, M., Fadel, L., & Saleh, G. (2018). Mental health in Syrian children with a focus on post-traumatic stress: a cross-sectional study from Syrian schools. *Social Psychiatry and Psychiatric Epidemiology*, 53(11), 1231–1239. <https://doi.org/10.1007/s00127-018-1573-3>
- Piper, B. (2010). *Ethiopia Early Grade Reading Assessment* (No. 7). Research Triangle Park, NC: RTI International.
- Platas, L. M., Ketterlin-Geller, L. R., & Sitabkhan, Y. (2016). Using an Assessment of Early Mathematical Knowledge and Skills to Inform Policy and Practice: Examples from the Early Grade Mathematics Assessment. *International Journal of Education in Mathematics, Science and Technology*, 4(3), 163–173.
- Pratham. (2013). *Annual Status of Education Report 2005-2012*. Mumbai: Pratham Resource Centre.
- RTI International. (2009). *Early Grade Reading Assessment Toolkit*. Retrieved from RTI International website: [http://www.ineesite.org/uploads/files/resources/EGRA{\\\\_}Toolkit{\\\\_}Mar09.pdf](http://www.ineesite.org/uploads/files/resources/EGRA{\\_}Toolkit{\\_}Mar09.pdf)
- Save the Children. (2019). Global Education Research. Retrieved June 16, 2019, from Save the Children website: <https://www.savethechildren.org/us/what-we-do/global-programs/education/research>
- Scales, P. C., Roehlkepartain, E. C., Wallace, T., Inselman, A., Stephenson, P., & Rodriguez, M. (2015). Brief report: Assessing youth well-being in global emergency settings: Early results from the Emergency Developmental Assets Profile. *Journal of Adolescence*, 45, 98–102. <https://doi.org/10.1016/j.adolescence.2015.09.002>
- Shaw, J. A. (2003). Children Exposed to War/Terrorism. *Clinical Child and Family Psychology Review*, 6(4), 237–246. <https://doi.org/10.1023/B:CCFP.0000006291.10180.bd>
- Sirin, S. R., & Rogers-Sirin, L. (2015). *The educational and mental health needs of Syrian refugee children*. Retrieved from Migration Policy Institute website: <http://www.migrationpolicy.org/research/educational-and-mental-health-needs-syrian-refugee-children>
- Spratt, J., King, S., & Bulat, J. (2013). *Independent evaluation of the effectiveness of Institut pour l'Education Populaire's Read-Learn-Lead (RLL) Program in Mali* [Technical report]. Research Triangle Park, NC: RTI International.
- Stemler, S. E., & Tsai, J. (2008). Best practices in interrater reliability: Three common approaches. In *Best practices in quantitative methods* (pp. 29–49). Thousand Oaks: SAGE.
- Tabachnick, B. G., Fidell, L. S., & Ullman, J. B. (2007). *Using multivariate statistics* (Vol. 5). Pearson Boston, MA.

- Thabet, A. A., Ibraheem, A. N., Shivram, R., Winter, E. A., & Vostanis, P. (2009). Parenting Support and PTSD in Children of a War Zone. *International Journal of Social Psychiatry*, 55(3), 226–237. <https://doi.org/10.1177/0020764008096100>
- The Economist. (2019, March 14). Syria's broken schools will make it difficult to fix the country. *The Economist*. Retrieved from <http://www.economist.com/middle-east-and-africa/2019/03/14/syrias-broken-schools-will-make-it-difficult-to-fix-the-country>
- The World Bank. (2018, July 1). Harmonized List of Fragile Situations. Retrieved June 15, 2019, from World Bank website: <http://www.worldbank.org/en/topic/fragilityconflictviolence/brief/harmonized-list-of-fragile-situations>
- Trani, J.-F., Fowler, P., Bakhshi, P., & Kumar, P. (2019). Assessment of progress in education for children and youth with disabilities in Afghanistan: A multilevel analysis of repeated cross-sectional surveys. *PLOS ONE*, 14(6), e0217677. <https://doi.org/10.1371/journal.pone.0217677>
- 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.
- UNESCO. (2011). *The hidden crisis: Armed conflict and education* [Education for All Global Monitoring Report]. Paris: UNESCO.
- UNESCO, & ALECSO. (2014). *Regional mapping report on assessment in the Arab States: Survey of student assessment systems in the Arab States*. Beirut, Lebanon: UNESCO Office Beirut.
- Wuermli, A. J., Tubbs, C. C., Petersen, A. C., & Aber, J. L. (2015). Children and Youth in Low- and Middle-Income Countries: Toward An Integrated Developmental and Intervention Science. *Child Development Perspectives*, 9(1), 61–66. <https://doi.org/10.1111/cdep.12108>
- Yoshikawa, H., Leyva, D., Snow, C. E., Treviño, E., Barata, M., Weiland, C., ... others. (2015). Experimental impacts of a teacher professional development program in Chile on preschool classroom quality and child outcomes. *Developmental Psychology*, 51(3), 309.
- Yoshikawa, H., & Way, N. (2008). From peers to policy: How broader social contexts influence the adaptation of children and youth in immigrant families. *New Directions for Child and Adolescent Development*, 2008(121), 1–8. <https://doi.org/10.1002/cd.219>
- Yousafzai, A. K., Rasheed, M. A., Rizvi, A., Armstrong, R., & Bhutta, Z. A. (2014). Effect of integrated responsive stimulation and nutrition interventions in the

Lady Health Worker programme in Pakistan on child development, growth, and health outcomes: a cluster-randomised factorial effectiveness trial. *The Lancet*, 384(9950), 1282–1293.

Appendix Table 1. *Subtask distribution assumptions*

<b>Total scores</b>	
Letter sound	Censored from above (50)
Familiar words	Censored from above (25)
Oral passage reading	Censored from below (0)
Reading comprehension	Categorical (7 categories)
Listening comprehension	Categorical (7 categories)
Vocab	Categorical (7 categories)
Number identification	Censored from above (20)
Number discrimination	Censored from above (10)
Missing number	Normal
Addition	Categorical (8 categories)
Subtraction	Categorical (7 categories)
Word problems	Categorical (7 categories)



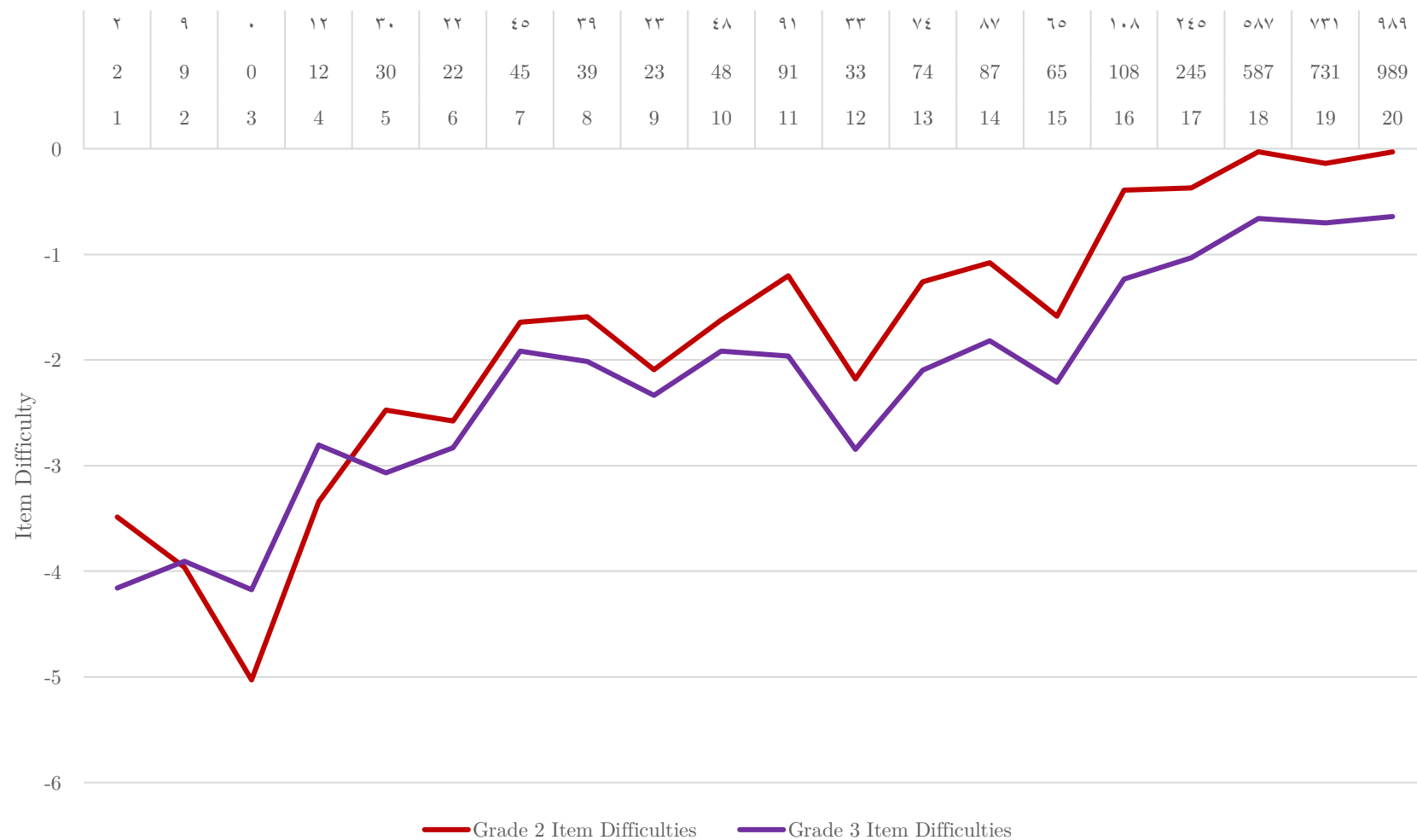
Appendix Table 2. *Item-level numeracy subtask unidimensional confirmatory model fit*

					RMSEA	
		X <sup>2</sup>	df	RMSEA	90% CI	CFI
nid_t	g2	611.50	170	0.059	0.054-0.064	0.965
	g3	454.74	170	0.048	0.043-0.054	0.954
dis_t	g2	187.60	35	0.077	0.066-0.088	0.944
	g3	95.99	35	0.049	0.038-0.061	0.946
miss_t	g2	359.06	35	0.129	0.117-0.141	0.864
	g3	270.69	35	0.112	0.100-0.125	0.878
add_t	g2	139.70	14	0.110	0.094-0.127	0.936
	g3	87.64	14	0.086	0.069-0.103	0.952
sub_t	g2	104.49	9	0.120	0.100-0.141	0.955
	g3	42.49	9	0.072	0.051-0.095	0.982

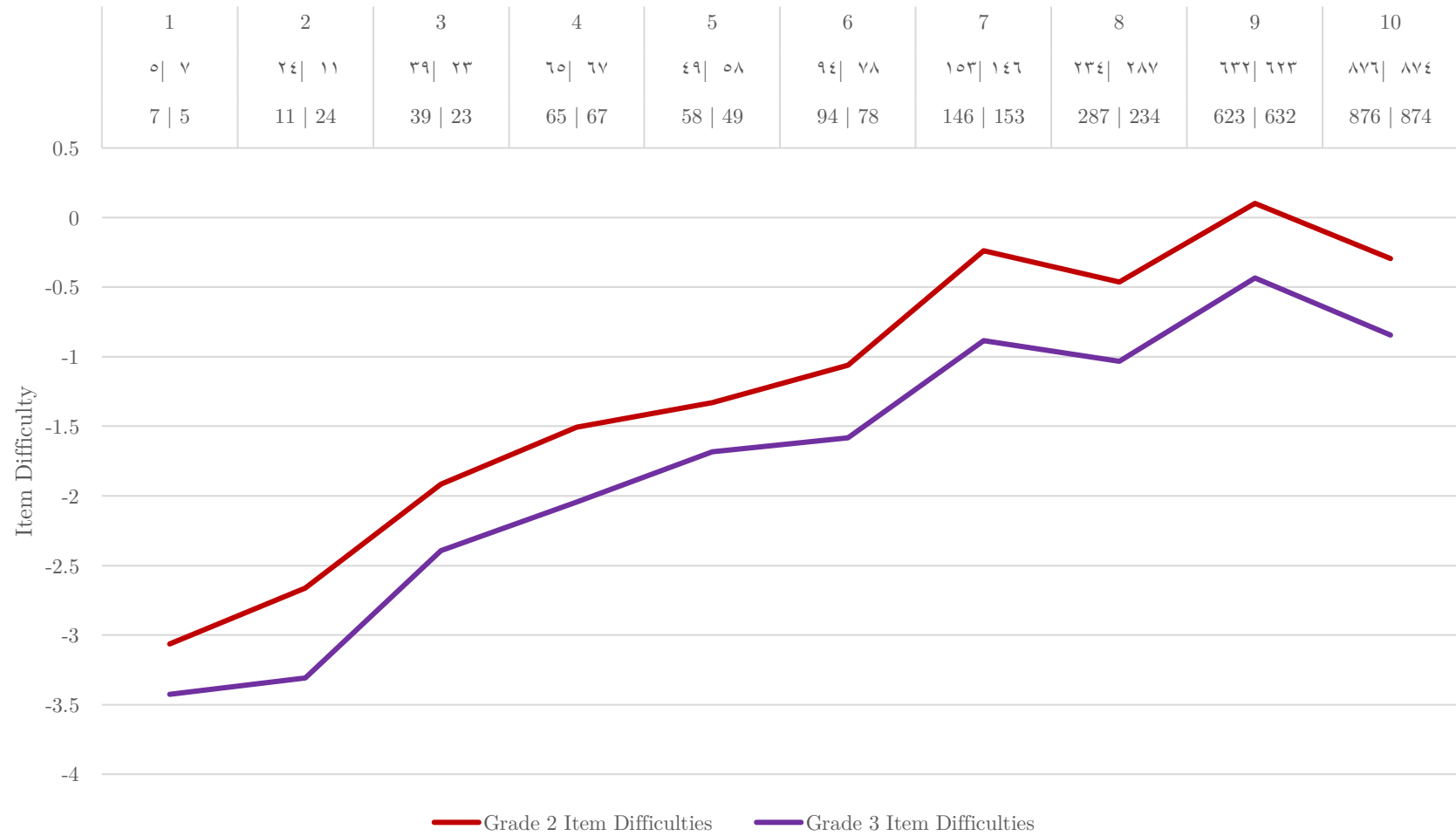
Appendix Table 2. *Item-level numeracy subtask IRT model comparisons*

		$X^2$	df	p	$\Delta AIC$	$\Delta BIC$
nid_t	g2	303.43	19	0	265.43	177.91
	g3	129.15	19	0	91.15	40.12
dis_t	g2	34.44	9	0	16.44	-25.02
	g3	10.12	9	0.341	-7.88	-48.98
add_t	g3	27.78	6	0	15.78	-11.60
sub_t	g3	16.77	5	0.005	6.76	-16.05

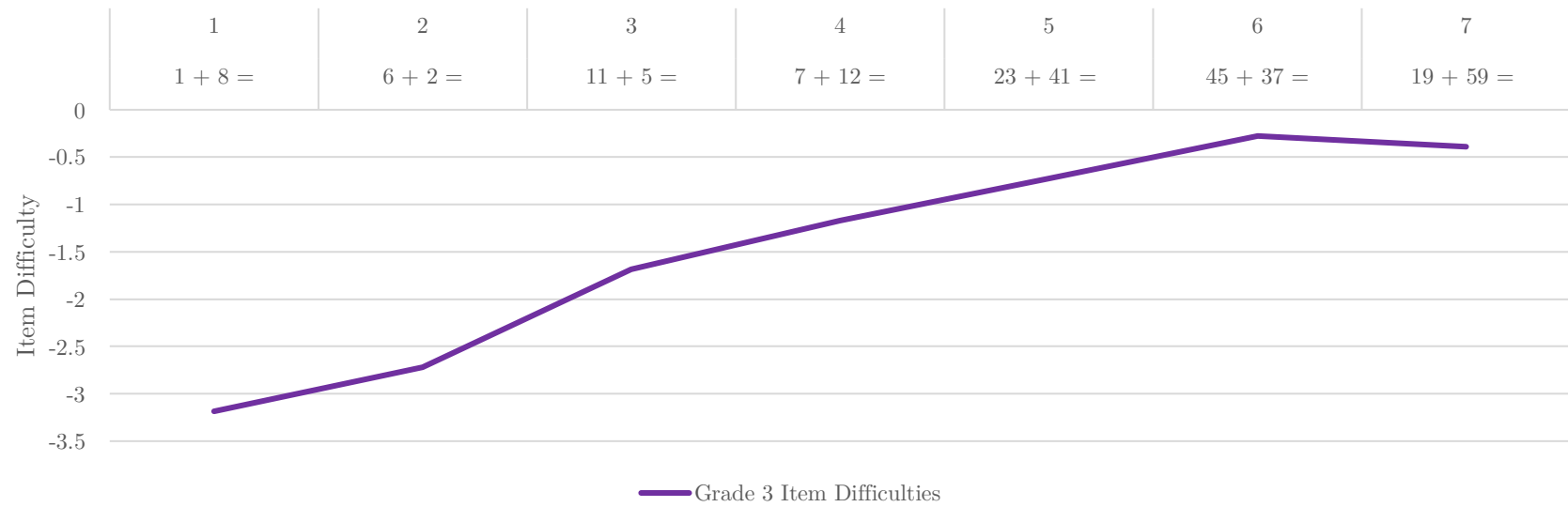
Appendix Figure 1. *Item difficulties (2-PL) for number identification subtask, by grade*



Appendix Figure 2. *Item difficulties (2-PL) for number discrimination subtask, by grade*



Appendix Figure 3. *Item difficulties (2-PL) for addition subtask, grade 3*



Appendix Figure 4. *Item difficulties (2-PL) for subtraction subtask, by grade*

