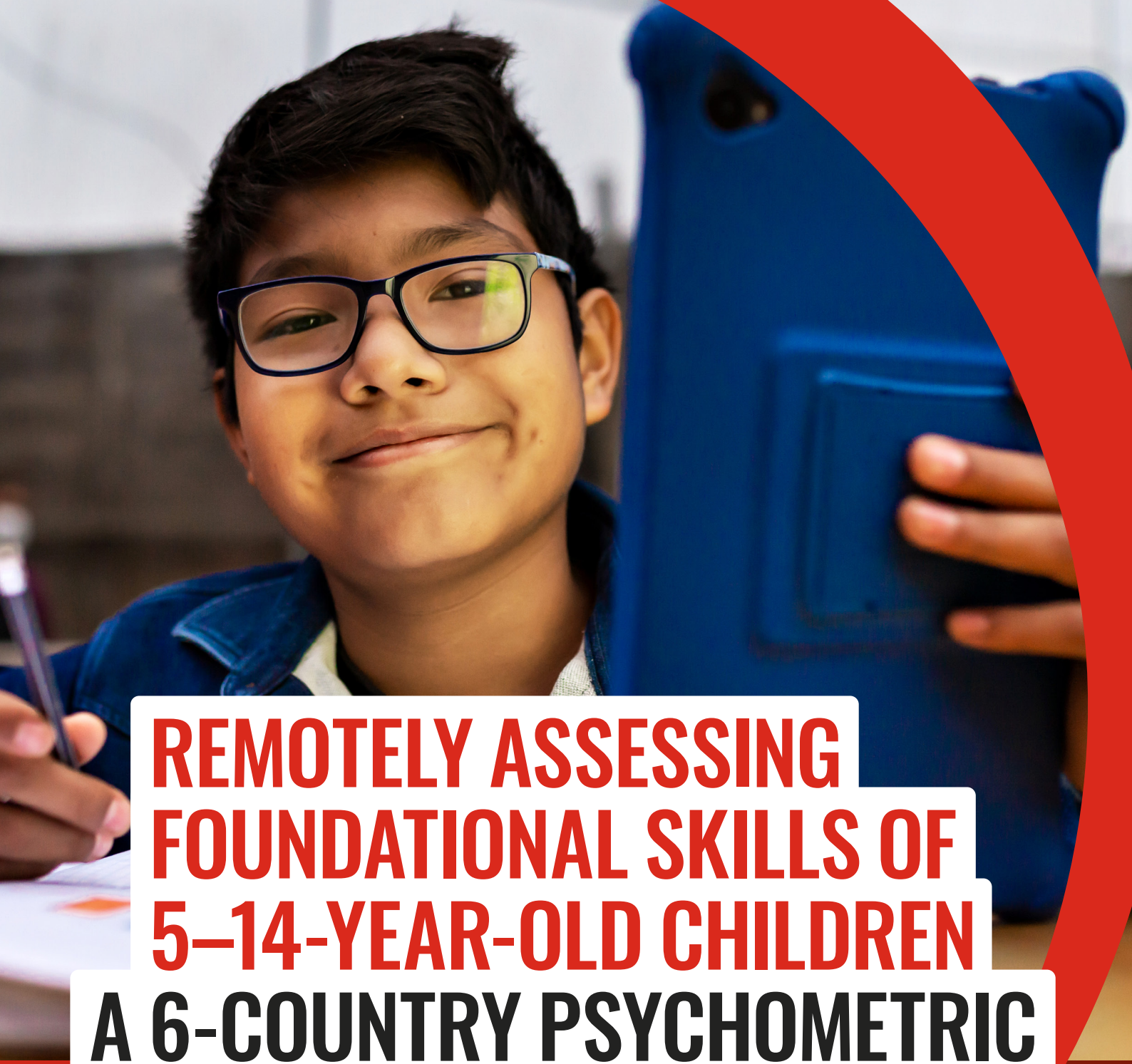


REMOTELY ASSESSING FOUNDATIONAL SKILLS OF 5–14-YEAR-OLD CHILDREN

A 6-COUNTRY PSYCHOMETRIC EVALUATION OF THE REMOTE ASSESSMENT OF LEARNING (ReAL)

YALE CHILD STUDY CENTER
Where discovery inspires care

Save the Children

AUTHORS

Elizabeth Hentschel
Yale Child Study Center (Yale University)
and
Sascha Hein
Freie Universität Berlin,
Yale Child Study Center (Yale University)

Nan Li
University of Texas Rio Grande Valley
Clay Westrope
Save the Children United States
Julia Taladay
Save the Children United States
Gillian Valentine
Save the Children International
James Leckman
Yale Child Study Center
Amira Abdurahman
Save the Children Sudan
Fatime Bachir
Save the Children Niger
Farah Darwazeh
Palestinian Child Institute
(An-Najah National University)
Xerxes de Castro
Save the Children Philippines
Doa Hamdan
An-Najah National University
Sithon Khun
Save the Children Cambodia
Sakem Kong
Save the Children Cambodia
Harouna Mounkaila
Abdou Moumouni University
Janet Mugo
Save the Children Sudan
Mohamed Sagayar
Abdou Moumouni University
Ali Nashat Shaar
An-Najah National University
Borin Srey
Save the Children Cambodia
Adriano S. Uaciquete
Universidade Eduardo Mondlane,
Ghent University
Megha Pande
Save the Children United States
Maria A. Gutierrez
Yale Child Study Center (Yale University)
Angelica Ponguta
Yale Child Study Center (Yale University)

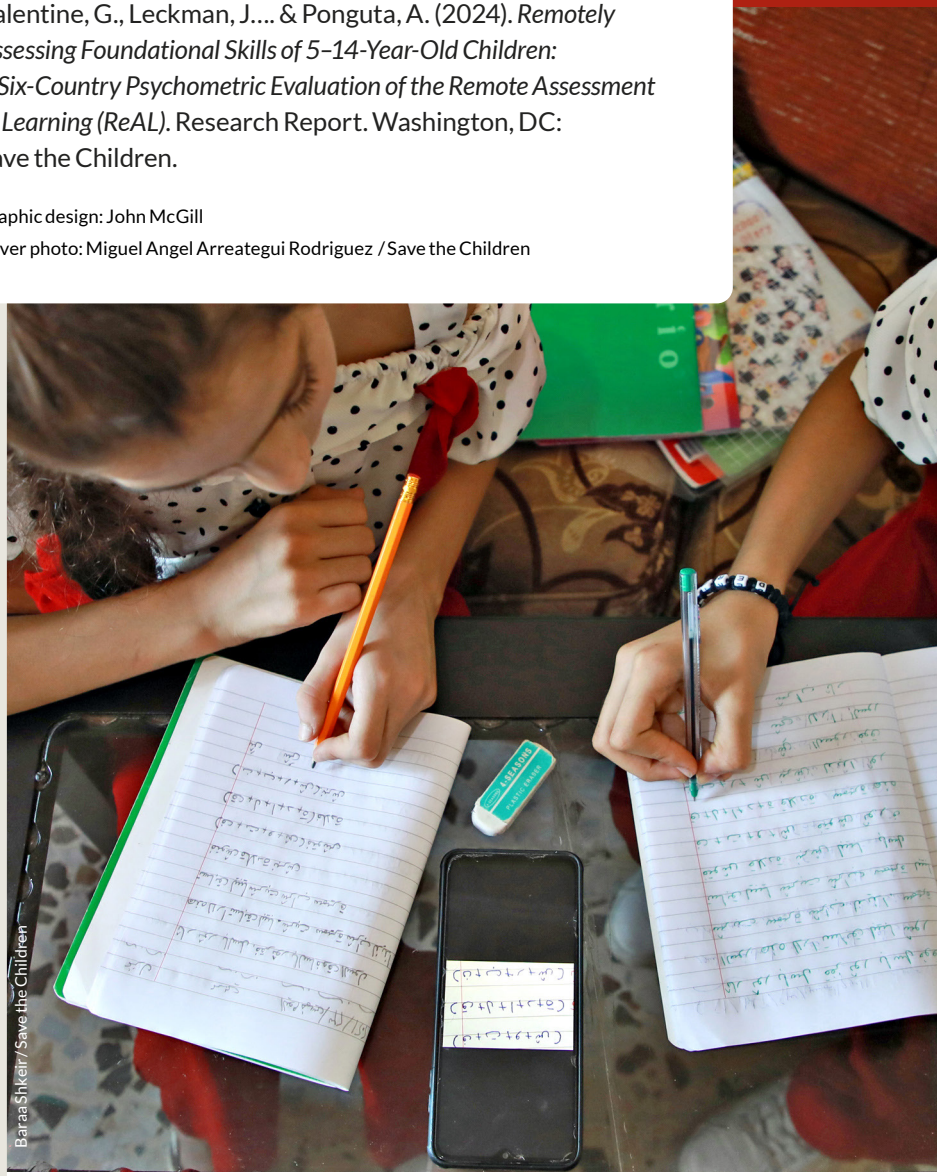
ACKNOWLEDGEMENTS

We would like to acknowledge the fruitful collaboration with members of the ReAL Network. This work would not have been possible without the support and partnership of local academic institutions and Save the Children country office teams, whose involvement has greatly enhanced the relevance and impact of this study. We also would like to thank the participants in this validation study – the children and their caregivers – for being willing to engage in this study and for giving their time to support the validation of the ReAL tool.

Suggested Citation

Hentschel, E., Hein, S., Li, N., Westrope, C., Taladay, J., Valentine, G., Leckman, J.... & Ponguta, A. (2024). *Remotely Assessing Foundational Skills of 5–14-Year-Old Children: A Six-Country Psychometric Evaluation of the Remote Assessment of Learning (ReAL)*. Research Report. Washington, DC: Save the Children.

Graphic design: John McGill
Cover photo: Miguel Angel Arreategui Rodriguez / Save the Children


Barasa Shkeir / Save the Children

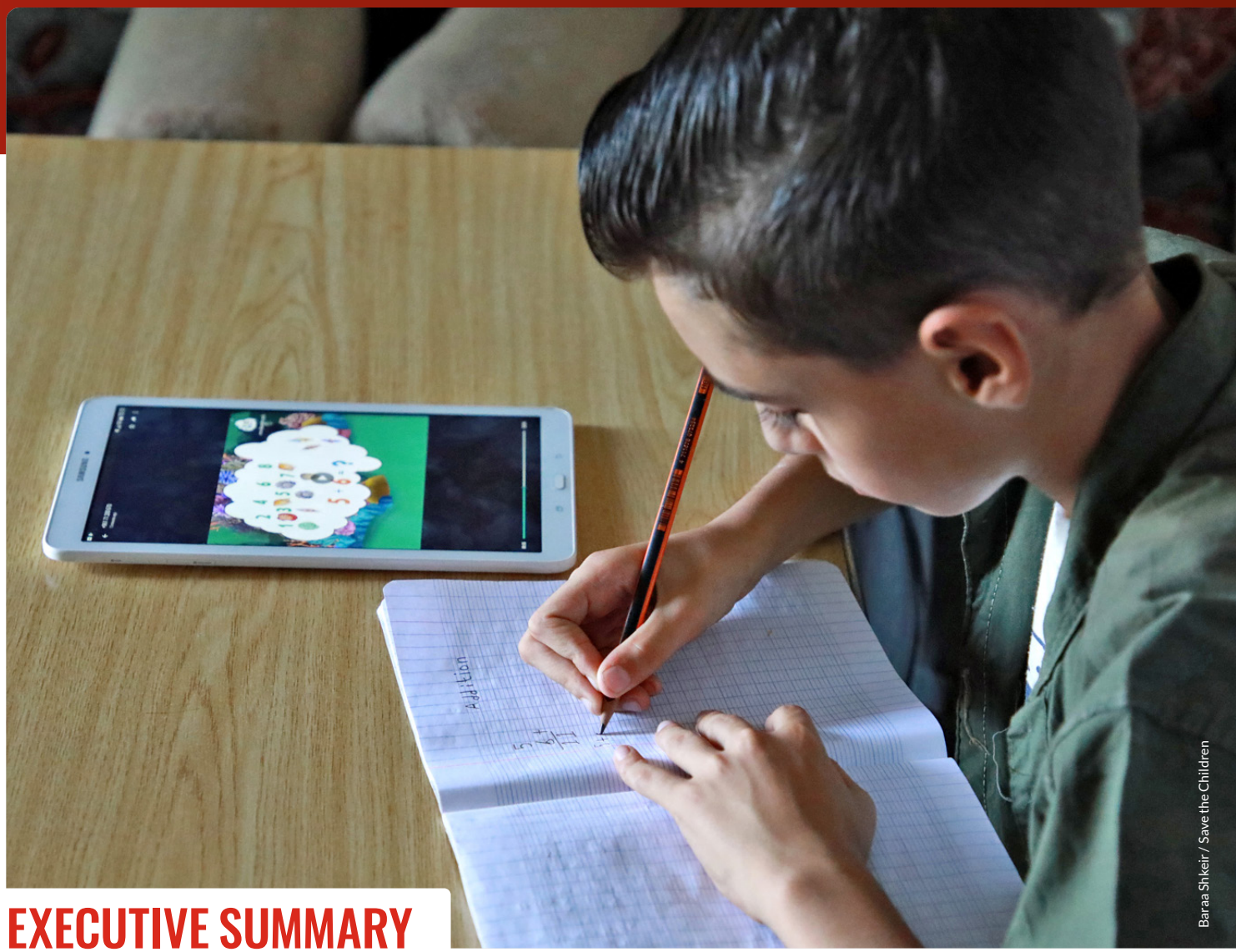
CONTENTS

ACRONYMS

CASEL	Collaborative for Academic, Social, and Emotional Learning
CFI	Comparative Fit Index
CFA	Confirmatory Factor Analysis
HALDO	Holistic Assessment of Learning and Development Outcomes
HICs	High-Income Countries
IDELA	International Development and Early Learning Assessment
IIF	Item Information Function
IRB	Institutional Review Board
IRR	Interrater Reliability
IRT	Item Response Theory
ISELA	International Social-Emotional Learning Assessment
LMICs	Low-and Middle-Income Countries
MAPS	Multidimensional Assessment of Parenting Scale
MEAL	Monitoring, Evaluation, Accountability, and Learning
N/A	Not Applicable
PSS	Psychosocial Support Scale
ReAL	Remote Assessment of Learning
RMSEA	Root Mean Square Error of Approximation
RTI	Research Triangle Institute
SD	Standard Deviation
SDG	Sustainable Development Goal
SEL	Social-Emotional Learning
SRMR	Standardized Root Mean Squared Residual
TIF	Test Information Function
TRT	Test-Retest Reliability
UN	United Nations

ABSTRACT

Approximately 250 million children worldwide are out of school. There is growing consensus for investing in feasible, contextually appropriate, psychometrically tested, remote tools to support quality education in crisis contexts. Save the Children developed the Remote Assessment of Learning to assess 5–14-year-old children's foundational learning. Children (N=4,840) were sampled from Cambodia, Mozambique, Niger, oPt, the Philippines, and Sudan, with an approximate 50/50 split in child sex within each country. The study assessed inter-rater reliability, factor structure, item slope and difficulty, criterion validity, and test-retest reliability. The study also explored user perceptions of feasibility and scalability. Results show moderate evidence that ReAL is valid and reliable for literacy and numeracy; evidence for social-emotional skills is weaker. The qualitative results revealed that while the tool is generally perceived as scalable and contextually appropriate, challenges persist with unreliable connectivity, caregiver influence, and comprehension issues in rural and linguistically diverse settings. This is the first cross-country evaluation of a remote assessment of learning.



Baraa Shkeir / Save the Children

EXECUTIVE SUMMARY

BACKGROUND

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 investments in the most educationally marginalized children: The approximately 250 million children worldwide who are out of school and the millions more not learning in school, a situation further exacerbated by conflict and crisis (UNESCO, 2023). The United Nations (UN) estimates that by 2030, 300 million students will lack the basic numeracy and literacy skills essential for full participation in today's world (United Nations, 2023). Importantly, global stakeholders increasingly converge on the importance of investing in field-feasible, contextually appropriate, and psychometrically sound measurement tools to support achieving quality education in crisis contexts (Tubbs Dolan, 2019).

These tools can provide accurate and timely data — what is often referred to as the "lifeblood" of the SDGs (Sachs, 2012, p. 2210) — about critical dimensions of children's learning and holistic development to support evidence-based decision-making within education systems. In the last two decades, widespread use of orally administered assessments of learning for pre-primary and primary school-aged children in low-and middle-income countries (LMICs) has helped policymakers and educators understand children's progress in reading, numeracy, and increasingly in social-emotional learning (SEL) (Montoya et al., 2016; Mulligan & Ayoub, 2023; Sowa et al., 2021). The development of sound assessments is an important scientific inquiry and a moral imperative to safeguard children's right to quality education. Such culturally relevant and scalable assessments are needed to better understand learning gaps of children – a need that is underscored by the high economic cost of the lack of formal education (UNESCO, 2024).

The development of the ReAL tool built upon an existing body of literature that has been working to conceptually and empirically understand child development from within unique cultural contexts and settings (e.g., Eisenberg et al., 2001; Oburu et al., 2016; Panter-Brick et al., 2017; Yoshikawa et al., 2008) and the Collaborative for Academic, Social, and Emotional Learning (CASEL) framework (CASEL, n.d.). One key aim was to infuse global policy, developmental, and measurement research with culturally appropriate developmental science insights (e.g., Barbot et al., 2020; Halpin et al., 2019; Jordans et al., 2013; Wuermli et al., 2015; Yoshikawa et al., 2015; Yousafzai et al., 2014). The ReAL Network, comprised of Save the Children, local academic partners, and a global consortium of thematic practitioners and researchers with psychometric expertise, contributed to the development of feasible, valid, reliable, and contextually appropriate measures for assessing foundational skill development of hard-to-reach children (e.g., children impacted by school closures, children not enrolled in school).

The approach was designed to address:

1. a lack of rigorous evidence on remotely administered assessments in LMICs (Angrist et al., 2022);
2. a gap in research-to-practice mechanisms in LMICs due to limited resources to translate research into policy or practice (Shumba et al., 2021); and
3. that most measures used in LMICs are developed by or drawn from educational science in HICs (e.g., Early Grade Reading Assessment) and the extent to which they provide valid and reliable data in crises has been called into question (Bartlett, et al., 2015; Dowd et al., 2019; Halpin & Torrente, 2014).

The ReAL tool built upon assessments such as the Holistic Assessment of Learning and Development Outcomes (HALDO; Krupar et al., 2019; Krupar & D'Sa, 2024), International Development and Early Learning Assessment (IDELA; Halpin et al., 2018; Pisani, et al., 2018; Wolf et al., 2017), International Social-emotional Learning Assessment (ISELA; D'Sa, 2019), Literacy Boost Reading Assessment, and Numeracy Boost Assessment to ensure that items were previously validated in face-to-face settings and that the assessment structure is familiar to implementers. The 5–14-year age range was selected for the pilot stage, as many of the instruments from which ReAL items are drawn had been used with children of these ages, and in many of the contexts in which this instrument would be used there are children in this age range learning foundational skills.

STUDY PURPOSE & RESEARCH QUESTIONS

The purpose of this study was to evaluate the psychometric properties of the versions of ReAL that country teams decided to test in their contexts (i.e., High Access and Caregiver Report) and understand user perceptions around feasibility and appropriateness for different operating contexts. The specific research questions are as follows:

1. **RQ1:** What are the psychometric properties of ReAL to remotely assess foundational skill development in seven LMICs?
   - a Assess the inter-rater reliability of the ReAL tool within each country;
   - b Identify the underlying factor structure of the ReAL tool across countries;
   - c Understand the criterion validity by measuring the extent to which the ReAL domains correlate with existing non-remote assessments of learning;
   - d Evaluate the test-retest reliability of the ReAL tool within each country; and
   - e Assess the practical relevance of the tool by understanding each item's difficulty.
2. **RQ2:** What are the perceptions of users of the ReAL tool during the validation study about its feasibility and appropriateness in different operating contexts?

Findings from this validation study will establish the extent to which a theoretically based remote assessment of learning is reliable and valid within and across seven LMICs. These insights will inform recommendations for future revisions of the assessment.

METHODOLOGY & LIMITATIONS

This study is comprised of a quantitative component examining the psychometric properties of the ReAL tool and a qualitative element exploring ReAL user perceptions of feasibility and appropriateness.

For the quantitative study, participants were sampled from Cambodia (n =1,108), El Salvador (n = 824), Mozambique (n = 458), Niger (n = 854), oPt (n = 1,135), the Philippines (n = 798), and Sudan (n = 587). The qualitative sampling strategy included a convenience sample of users engaged in using ReAL.

Three types of measures were used: (1) the ReAL tool, (2) a battery of criterion measures selected by each country team, and (3) qualitative protocols. Below we present the quantitative analyses conducted for each country in the sample.

Country	Descriptive	IRR	CFA	Criterion	TRT	IRT
Cambodia	×	×	×	×	×	×
El Salvador	×	×	×	×	×	×
Mozambique	×	×	×	×	×	×
Niger	×	×	×	×	×	×
oPt	×	×	×	×	×	×
Philippines	×	×	×	×	×	×
Sudan	×	×	×			×

The qualitative data were analyzed through a systematic manual process, with pre-determined themes of feasibility and appropriateness guiding the analysis. Initially, responses were organized by question, grouping the answers from each office together to maintain a coherent structure for analysis. The data were then carefully reviewed, with responses manually sorted into the pre-determined categories of feasibility and appropriateness, along with their corresponding sub-themes. This was achieved by highlighting key points. A summary of the findings was then compiled, focusing on the frequency and significance of the identified sub-themes across different contexts. This process allowed for a comprehensive understanding of user perceptions.

FINDINGS

The primary goal of the current study was to assess the psychometric properties of the High Access and Caregiver Report modalities of ReAL. Specifically, we appraised the inter-rater reliability, underlying factor structure, criterion validity, test-retest reliability, and item slope and difficulty. Our results show, with only a few notable exceptions, moderate to strong evidence that ReAL is a valid and reliable measure of the literacy and numeracy sub-domains assessed. The evidence for the social-emotional sub-domains is less robust or lacking.

A secondary objective of this study was to explore the perceived feasibility and appropriateness of using ReAL in different operating contexts. Specifically, we evaluated perceptions of feasibility by examining factors such as technical infrastructure, logistical challenges, the assessment environment, staff training, and scalability. We also assessed perceptions of appropriateness through considerations of contextualization, content relevance, and suitability. The qualitative findings revealed that while the tool is generally perceived as scalable and contextually appropriate, challenges persist with unreliable connectivity, caregiver influence, and comprehension issues in rural and linguistically diverse settings. Additionally, concerns were raised about the tool's relevance to the assessed grade levels indicating a need for further contextual adaptation.

DISCUSSION

Through this study, we demonstrate that a remote assessment of learning can be a feasible, valid, and reliable measure of foundational academic (i.e., literacy and numeracy) skills in LMICs. While this evidence is promising, there are critical preconditions that must be met when conducting a remote, phone-based assessment: A cellular network infrastructure and connectivity and ownership of or access to a phone. These conditions may not always be present in many low-resource contexts. For this reason, it is important to understand the context in which the assessment will take place to evaluate whether such an assessment is appropriate. This will also aid in identifying the linguistic background of the participants, allowing for the tools to be accurately translated and adapted to the specific language needs of the context. Thus, we advocate for assessments like ReAL to be added to the other options educations systems have to assess foundational skills rather than serving as a replacement. We also identified one other precondition for the successful implementation of the tool: constructive caregiver support. Addressing the role of caregivers during assessments may involve creating guidelines or training materials to ensure that this involvement is guided and managed appropriately, enhancing rather than detracting from the accuracy and reliability of the data collected.

Given the promise of remote, phone-based assessments as one option to assess foundational skills of hard-to-reach children in LMIC contexts, we advocate for further research that builds upon this validation study. Specifically, we propose developing and testing adaptive versions that could more efficiently assess skills across the wide age range we target here and piloting the tool in more diverse settings to refine its reach and effectiveness, particularly among vulnerable populations in rural areas Additionally, building evidence of cost effectiveness in comparison with other assessment types will be a critical consideration for any education system weighing a phone-based assessment like ReAL with one administered face-to-face. Finally, we seek to make the literacy and numeracy items more difficult and to re-develop and test the ReAL SEL sub-domains in future studies.

Patimiosi Ngungutau/ Save the Children

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 investments in the most educationally marginalized children: The approximately 250 million children worldwide who are out of school and the millions more not learning in school; an issue further exacerbated by conflict and crisis (UNESCO, 2023). The United Nations (UN) estimates that by 2030, 300 million students will lack the basic numeracy and literacy skills essential for full participation in today's world (United Nations, 2023). Importantly, global stakeholders increasingly converge on the importance of investing in field-feasible, contextually appropriate, and psychometrically sound measurement tools to support achieving quality education in crisis contexts (Tubbs Dolan, 2019).

These tools can provide accurate and timely data — what is often referred to as the "lifeblood" of the SDGs (Sachs, 2012, p. 2210) — about critical dimensions of children's learning and holistic development to support evidence-based decision-making within education systems. In the last two decades, widespread use of orally administered assessments of learning for pre-primary and primary school-aged children in low-and middle-income countries (LMICs) has helped policymakers and educators understand children's progress in reading, numeracy, and increasingly in social-emotional learning (SEL) (Montoya et al., 2016; Mulligan & Ayoub, 2023; Sowa et al., 2021). The development of sound assessments is an important scientific inquiry and a moral imperative to safeguard children's right to quality education. Such culturally relevant and scalable assessments are needed to better understand learning gaps of children – a need that is underscored by the high economic cost of the lack of formal education (UNESCO, 2024).

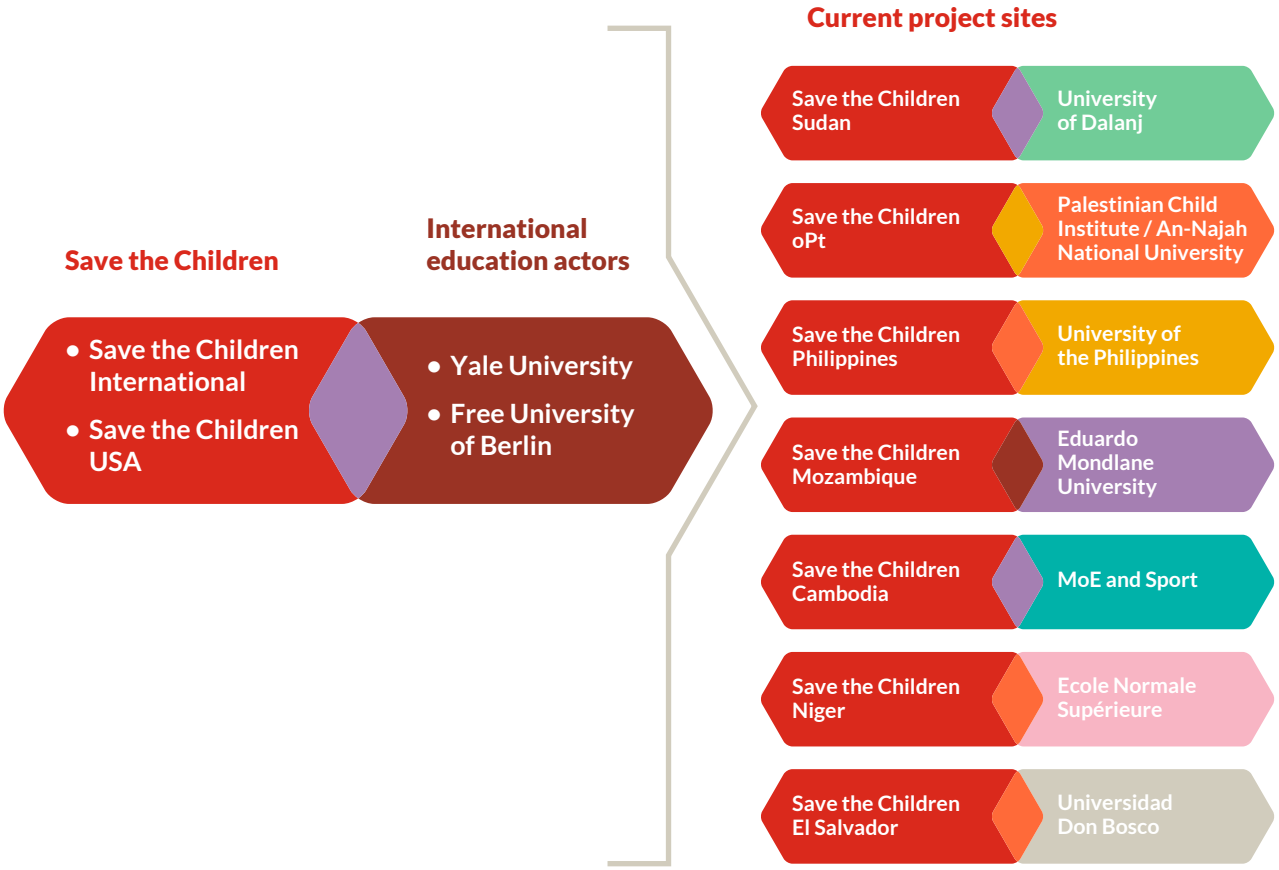
While there has been growth in development and testing of learning assessments administered face-to-face, there is limited evidence for learning assessments administered remotely. There is an urgent need to develop and test such assessments as educators, officials and humanitarian actors in crisis-affected settings have consistently faced the challenge of how to assess children's academic outcomes and social-emotional skills when lockdowns, school closures and other unexpected crises prevent the use of face-to-face assessment tools. The 2020 COVID-19 pandemic exacerbated and highlighted this challenge, as 214 million students from pre-primary to upper secondary education in 23 countries missed at least three quarters of classroom instruction time over one academic year due to school closures (UNICEF, 2021). Distance learning programs were implemented, but practitioners lacked valid, reliable, relevant and feasible remote assessment tools to evaluate students' growth of academic and social-emotional skills when engaging with these programs. Most of the existing remote learning assessment of learning tools have been created for the adult or elder learning space (Rapp et al., 2012; Sticht et al., 1996), or use a hybrid model that includes both in-person and remote assessment (Aker & Ksoll, 2020), but the feasibility, reliability and validity of the existing remote assessments in this space are promising. To the best of our knowledge, there are only two examples of fully remote, phone-based assessments that have been used to assess learning in LMICs: a numeracy assessment that was developed by Angrist et al. (2022) and tested in Botswana among primary aged children that was found to accurately capture basic numeracy skills, and a language and literacy assessment conducted by Sobers et al. (2023) in Cote d'Ivoire that was found to be valid and reliable. In response to this gap in the literature and programmatic need, Save the Children developed the Remote Assessment of Learning (ReAL) tool to remotely assess literacy, numeracy, and social-emotional outcomes for children ages 5–14-years-old.

The development of the ReAL tool built upon an existing body of literature that has been working to conceptually and empirically understand child development from within unique cultural contexts and settings (e.g., Eisenberg et al., 2001; Oburu et al., 2016; Panter-Brick et al., 2017; Yoshikawa et al., 2008) and the Collaborative for Academic, Social, and Emotional Learning (CASEL) framework (CASEL, n.d.). One key aim was to infuse global policy, developmental, and measurement research with culturally appropriate developmental science insights (e.g., Barbot et al., 2020; Halpin et al., 2019; Jordans et al., 2013; Wuermli et al., 2015; Yoshikawa et al., 2015; Yousafzai et al., 2014).

The ReAL Network, comprised of Save the Children, local academic partners, and a global consortium of thematic practitioners and researchers with psychometric expertise, contributed to the development of feasible, valid, reliable, and contextually appropriate measures for assessing foundational skill development of hard-to-reach children (e.g., children impacted by school closures, children not enrolled in school). The approach was designed to address: (1) a lack of rigorous evidence on remotely administered assessments in LMICs (Angrist et al., 2022); (2) a gap in research-to-practice mechanisms in LMICs due to limited resources to translate research into policy or practice (Shumba et al., 2021); and (3) that most measures used in LMICs are developed by or drawn from educational science in HICs (e.g., Early Grade Reading Assessment) and the extent to which they provide valid and reliable data in crises has been called into question (Bartlett, et al., 2015; Dowd et al., 2019; Halpin & Torrente, 2014). The ReAL tool built upon assessments such as the Holistic Assessment of Learning and Development Outcomes (HALDO; Krupar et al., 2019; Krupar & D'Sa, 2024), International Development and Early Learning Assessment (IDELA; Halpin et al., 2018; Pisani, et al., 2018; Wolf et al., 2017), International Social-emotional Learning Assessment (ISELA; D'Sa, 2019), Literacy Boost Reading Assessment, and Numeracy Boost Assessment to ensure that items were previously validated in face-to-face settings and that the assessment structure is familiar to implementers. The 5–14-year age range was used at the pilot stage, as many of the instruments from which ReAL items are drawn had been used with children of these ages, and in many of the contexts in which this instrument would be used there are children in this age range learning foundational skills.

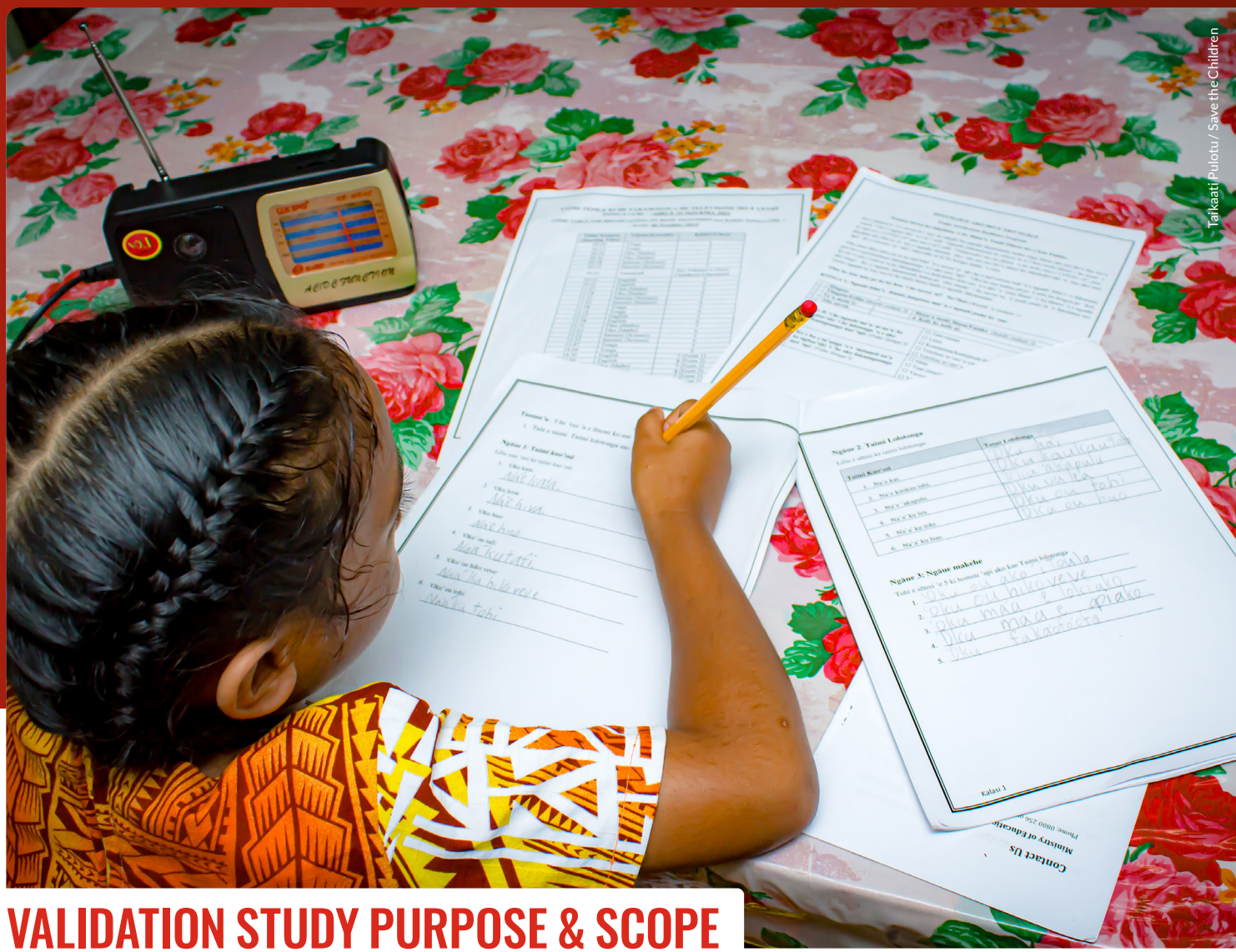
After the first iteration of the tool was developed, the ReAL was piloted in Bangladesh, Guatemala, the Philippines, and Zambia. This first pilot, known as the alpha phase, found high internal consistency reliability within the literacy (minimum alpha=0.86), numeracy (minimum alpha=0.78), and social-emotional learning (minimum alpha=0.90) domains. This indicated that the items provided a reliable assessment of the hypothesized domains and sub-domains in pilot sites. The alpha pilot identified some sub-domains that did not function well psychometrically or were difficult to administer based on feedback from pilot sites. These sub-domains and items were either removed or revised. Given the initial evidence that the ReAL had the potential to be a reliable and valid measure of these learning skills, the team embarked on the beta phase using the revised tool, which involved a rigorous validation study in seven countries with hard-to-reach populations: Cambodia, El Salvador, Mozambique, Niger, the Occupied Palestinian Territory (oPt), the Philippines, and Sudan. All countries, except El Salvador, selected the High Access version of ReAL. El Salvador selected the Caregiver Report version.

Figure 1
ReAL Network



This study was supported through collaboration with colleagues in Save the Children country offices and local academic partners as part of the ReAL Network, shown in Figure 1. Within the ReAL Network, local academic institutions are co-equal partners. They ensured that the ReAL tool was contextually relevant and culturally appropriate. The partnership in each country led the interpretation of the results from a local lens. By engaging local institutions at the validation stage, there is a greater likelihood of ReAL being used to inform education finance and planning once it is launched as a global public good and that local stakeholders have ownership of the results. This approach also ensured that data collection methods were embedded locally and that the measure is culturally valid and reliable.

The COVID-19 pandemic forced the global community to face a persistent gap that has been present in our foundational skills assessment toolbox and thus continued to limit our ability to deliver on the promise of SDG 4: our inability to assess hard-to-reach children when face-to-face access was not possible. The pandemic catalyzed the development and testing of phone-based, remote assessments in LMICs, providing us with some promising options. While the evidence remains thin, we are beginning to coalesce around some promising examples. These phone-based assessments of foundational learning skills have the potential to expand education systems' knowledge of the skill levels of even the hardest-to-reach children, providing critical information for education decision-makers. This study contributes to the evidence base on valid and reliable measures of foundational academic skills for children 5–14 years old and provides for future directions in the development of valid and reliable measures of foundational SEL skills.



Tarkaati Pulotu / Save the Children

VALIDATION STUDY PURPOSE & SCOPE

The purpose of this study was to evaluate the psychometric properties of the versions of ReAL that country teams decided to test in their contexts (i.e., High Access and Caregiver Report) and understand user perceptions around feasibility and appropriateness for different operating contexts. The specific research questions are as follows:

1 **RQ1:** What are the psychometric properties of ReAL to remotely assess foundational skill development in seven LMICs?

a Assess the inter-rater reliability of the ReAL tool within each country;

b Identify the underlying factor structure of the ReAL tool across countries;

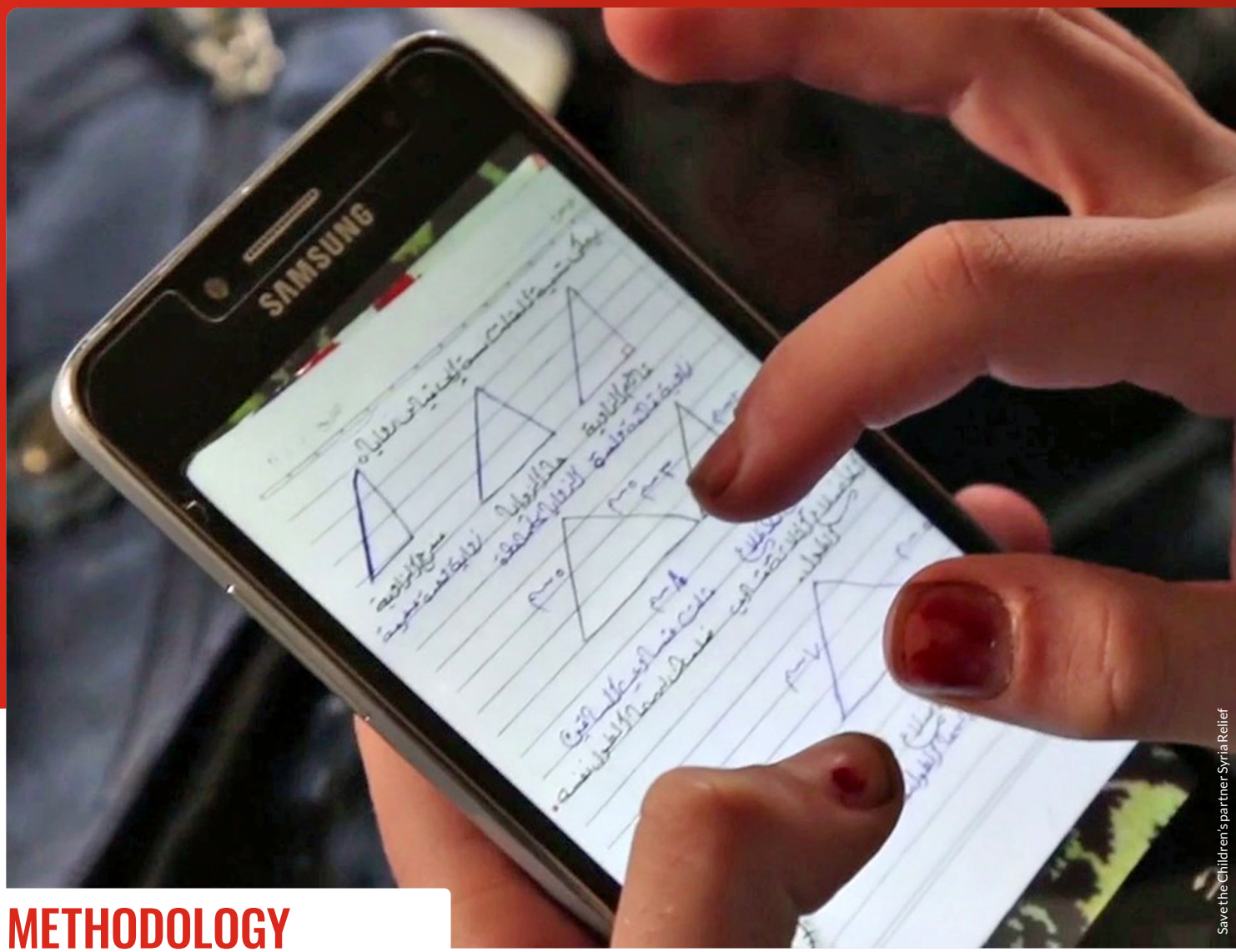
c Understand the criterion validity by measuring the extent to which the ReAL domains correlate with existing non-remote assessments of learning;

d Evaluate the test-retest reliability of the ReAL tool within each country; and

e Assess the practical relevance of the tool by understanding each item's difficulty.

2 **RQ2:** What are the perceptions of users of the ReAL tool during the validation study about its feasibility and appropriateness in different operating contexts?

Findings from this validation study establish the extent to which a theoretically based remote assessment of learning is reliable and valid within and across seven LMICs and allows us to make recommendations for item revisions for future iterations of the assessment.



Save the Children's partner Syria Relief

METHODOLOGY

In this section we present the sample and sampling strategy, the data sources, the approach to data analysis, the limitations of the study, and the ethics and accountability measures for this research. Each study was co-designed with country-level academic and practitioner partners through a structured curriculum and design process.

STUDY DESIGN & SAMPLING

This study is comprised of a quantitative component examining the psychometric properties of the ReAL tool and a qualitative element exploring ReAL user perceptions of feasibility and appropriateness.

The quantitative sampling strategy included a target sample size of 200-child caregiver dyads per selected age group in each country. The sample size was determined based on the statistical power required to conduct confirmatory factor analysis (CFA). In each site, the research team selected the most appropriate age groups to include in the validation study between the ages of 5 to 14-years-old. Teams sought equal representation of girls and boys in the sample. From this overall sample, a total of 40% of the children were to be randomly selected to be assessed: (1) by two enumerators to probe inter-rater reliability; (2) a second time four to 12 weeks after the first assessment to determine retest reliability; and (3) using external measures to examine criterion validity. The interrater reliability and test-retest sub-sample groups were to be mutually exclusive. The qualitative sampling strategy included a convenience sample of users engaged in using ReAL.

Participants were sampled from Cambodia (n =1,108), El Salvador (n = 824), Mozambique (n = 458), Niger (n = 854), oPt (n = 1,135), the Philippines (n = 798), and Sudan (n = 587). Table 1 below shows the sample size, age and standard deviation (SD), and sex distribution by country.

Table 1
ReAL Validation Study Sample Characteristics

Country	Total sample size	Age distribution						Sex distribution	
		5–6	7–8	9–10	11–12	13–14+	Average (SD)	Girls (%)	Boys (%)
Cambodia	1,108	234	252	254	258	11	9.15 (2.25)	529 (52.43%)	480 (47.57%)
El Salvador	824	183	158	169	169	145	9.84 (2.80)	405 (49.15%)	419 (50.85%)
Mozambique	458	0	9	66	112	271	13.23 (2.06)	216 (47.16%)	242 (52.84%)
Niger	854	28	147	249	296	133	10.75 (2.13)	450 (52.69%)	403 (47.19%)
Opt	1,135	139	258	311	288	147	10.04 (2.38)	577 (50.53%)	565 (49.47%)
Philippines	798	0	4	175	263	356	12.68 (1.80)	431 (54.01%)	367 (45.99%)
Sudan	587	27	87	147	169	194	11.44 (2.62)	365 (65.30%)	194 (34.70%)

DATA SOURCES

We present below the data sources used in this validation study, namely: (1) the ReAL tool, (2) a battery of criterion measures selected by each country team, and (3) qualitative protocols.

The ReAL

The ReAL tool is a first-of-its-kind remote assessment of learning developed to measure foundational skills of hard-to-reach children aged 5 to 14-years-old. It is intended to measure children's literacy, numeracy, and social-emotional learning (SEL) skills for program monitoring and evaluation purposes. The pilot version of the ReAL literacy domain contains an oral language module (with expressive vocabulary, listening comprehension and retelling a story sub-domains) and a reading module (with letter/letter sound identification, common word identification, sentence-level comprehension and oral passage reading sub-domains). Each sub-domain contains between 2–10 questions that are typically dichotomously scored as correct or incorrect, except for expressive vocabulary, which is a continuous response. The pilot version of the ReAL numeracy domain contains five sub-domains (one- to- one correspondence, number identification, addition, subtraction, and word problems). Each sub-domain contains between 3–10 questions and all answers are dichotomously scored as correct or incorrect. The ReAL SEL domain contains six sub-domains (self-concept, relationships, perseverance, stress management, empathy, and conflict resolution). Each sub-domain has between four and 11 questions and each answer is dichotomously scored as correct or incorrect, or as appropriate or inappropriate.

The ReAL can be administered over the phone in three different administration modalities (High Access, Low Access, or Caregiver-reported) depending on the extent of child accessibility, materials (stimuli) availability, and type of phone (smart or conventional) that the parents/caregivers have access to. This ensures that ReAL is accessible to everyone, regardless of their technological resources (Save the Children, n.d.). Figure 2 below summarizes the approach for each modality. The two tested modalities can be found in Appendix F (High Access) and Appendix G (Caregiver Report).

All countries in this validation study opted for the High Access modality, with the exception of El Salvador, which opted for the Caregiver Report version. The ReAL tool was translated into different languages: Arabic for Sudan and oPt, Portuguese for Mozambique, Tagalog for the Philippines, Khmer for Cambodia, French for Niger, and Spanish for El Salvador. Additional contextualization of the tool was conducted by practitioners and researchers in each site to ensure literacy items were appropriately levelled and SEL items included contextually appropriate and inappropriate response options. Following contextualization and training, assessors verbally guided the caregiver and child through a detailed protocol over the phone while assessment stimuli were shared via the phone using SMS, WhatsApp, or hard copy handed out to the caregivers prior to the assessment.

Figure 2
ReAL Modalities

High Access
Direct child assessment

- **If:** Children can access smart phone/WhatsApp OR audio phone with hard copy material
- **Then:** Administer assessment via phone and WhatsApp with stimuli sent via phone OR via phone on audio with hard copy stimuli in front of the child
- **1:** Smart phone + WhatsApp (stimuli sent via the app)
  2: Audio phone + hard copy material

Low Access
Direct child assessment

- **If:** Children can access audio phone BUT cannot access specific stimuli for the assessment AND do not have a smart phone
- **Then:** Administer via phone and use existing learning materials at home for assessment

Caregiver
Caregiver report only

- **If:** Caregivers have phones AND we cannot speak to their child
- **Then:** Ask caregivers to report on their child's learning skills

Criterion Measures

The criterion measures varied by country. In Mozambique, Niger, oPt, and the Philippines, the Early Grade Reading Assessment (RTI International, 2015) and the Early Grade Math Assessment (RTI International, 2014) were used to measure literacy and numeracy skills. For SEL, an adapted version of the Psychosocial Support Scale (PSS), previously validated in Sudan (Olayemi et al., 2021), was used as a criterion measure in Mozambique, Niger, oPt, and the Philippines. In Cambodia, a Ministry of Education tool was used as a criterion measure for all constructs. In El Salvador, a combination of three tools were used as criterion measures: (1) Thinker: Tool for mothers, fathers and caregivers, 2) Multidimensional Assessment of Parenting Scale (MAPS), and (3) A home learning environment measure.

Qualitative Protocols

The qualitative approach involved developing a key informant interview questionnaire aimed at exploring the perceived feasibility and appropriateness of using the ReAL tool across different operating contexts. Sub-themes were identified through a literature review and preliminary analysis of enumerator comments received during validation study, which were then categorized under feasibility and appropriateness. A total of 13 targeted questions were designed based on these sub-themes and distributed via email to participants in all seven countries involved in the validation study. This protocol can be found in Appendix H.

Additionally, enumerator comments collected through Kobo during the validation study were compiled, translated, cleaned, and coded for analysis. This feedback was disaggregated by country, and comments relevant to the themes of feasibility and appropriateness were included in the analysis. The protocol ensured a systematic approach to gathering and interpreting qualitative data, with ethical considerations such as participant confidentiality and informed consent being maintained throughout the process.

DATA ANALYSIS

Quantitative

Analyses were conducted in Stata v18 (Stata- Corp, College Station, Texas, USA) and RStudio (R Core Team, 2021). Descriptively, we analyzed the age distribution, grade distribution, languages spoken at home, relationship of caregiver to child, and caregiver sex by country. Descriptive statistics can be found in Appendix A. Spearman correlation coefficients were then calculated between each ReAL sub-domain.

Inter-rater reliability was calculated as the percent agreement between two independent enumerators and we used Graham et al.'s (2012) benchmarks to classify percent agreement. Spearman correlations were computed at the sub-domain level for test-retest reliability. For test-retest correlations, we used Cicchetti's (1994) defined recommendations for appraising the magnitude of correlations ranging between .40–.59 as fair, .60–.74 as good and above .75 as excellent. For both the inter-rater reliability and test-retest analyses, missing values among individuals with partial missingness were assumed to be incorrect. Individuals missing all information on a given sub-domain were dropped from that sub-domain analysis but included in other analyses where they had at least partial missingness.

We used the lavaan package (Version 0.6-17; Rosseel, 2012) embedded in the R environment to perform confirmatory factor analysis (CFA) for each sub-domain based on the a-priori, theoretical mapping of the ReAL items onto the literacy, numeracy, and SEL domains. Specifically, a one-factor CFA model was conducted separately for listening comprehension (5 items), letter/letter sound identification (20 items), common word identification (10 items), sentence-level comprehension (5 items), oral passage reading comprehension (7 items), number identification (12 items), addition (10 items), subtraction (10 items), self-concept (6 items), use of social supports (4 items), and help seeking behavior (4 items). A two-factor CFA model was developed separately for stress management (4 items loaded on social supports and 3 different items loaded on behavioral regulation), empathy (5 items loaded on identifying feelings of others and 6 items loaded on empathy), and conflict resolution (4 items loaded on social problem solving and 2 items loaded on interpreting hostility). CFA was not conducted for expressive vocabulary (2 items), retelling a story (2 items), one-to-one correspondence (3 items), and word problems (3 items) since a one-factor CFA model with two indicators was under-identified and a one-factor CFA model with three indicators was just-identified. We used the weighted least square mean and variance estimator (WLSMV) to generate accurate inferences for binary indicators.

Pairwise deletion was used to deal with missing data. Global goodness of fit was evaluated based on Chi-Square ($\chi 2$), Comparative Fit Index (CFI), Root Mean Square Error of Approximation (RMSEA) and its 90% confidence interval, and Standardized Root Mean Squared Residual (SRMR). Given that conventional CFI and RMSEA tend to be overestimated when indicators are ordered-categorical variables, we reported robust CFI and RMSEA proposed by Savalei (2021) when available. Values greater than 0.90 and 0.95 for both the CFI and TLI, respectively, are considered to reflect acceptable and excellent fit to the data, whereas values for RMSEA less than 0.05 and 0.08 indicate excellent fit and acceptable fit to the data, respectively (Hu & Bentler, 1999; Marsh et al., 2005). A SRMR value less than .08 was recommended (Hu & Bentler, 1999). Moreover, local model fit was evaluated by examining Bentler-type correlation residuals (type = "cor.bentler" in lavaan), which are reported in Appendix C for each CFA model.

For sub-domains of literacy and numeracy with empirical evidence supporting unidimensionality, we proceeded with appraising item characteristics by conducting item response theory (IRT) analysis. IRT was not conducted for SEL because the item parameters for SEL are not as informative as those for literacy and numeracy. Unlike literacy and numeracy that are often more clearly defined and objective, SEL constructs can be more subjective and complex, involving attitudes, behaviors, and emotions. These constructs might not fit well into the IRT framework, which assumes unidimensionality. Pragmatically, knowing the slope and the difficulty for each SEL item, for example, "Can you describe what you hope will happen in your life in the future" (item sel 1), does not provide valuable information for refining the item. As a result, they do not provide valuable information for refining the test. 2PL IRT model characterizing item slope and difficulty was carried out using the mirt package (Version 1.41; Chalmers, 2012) embedded in the R environment. Due to the relatively large number of sub-domains under examination and the quantity of items involved, we presented item information function (IIF) and test information function (TIF) instead of presenting parameter estimates. This is because IIF and TIF offer straightforward ways to visualize item- and test-level performance. The IIF and TIF graphs are presented in Appendices D and E.

Criterion validity was assessed by calculating Spearman correlation coefficients between the ReAL domains with their corresponding criterion measures for relevant literacy, numeracy and social emotional learning sub-domains.

We show in Table 2 below the specific analyses conducted for each country.

Table 2
Quantitative Analyses Conducted by Country

Country	Descriptive	IRR	CFA	Criterion	TRT	IRT
Cambodia	×	×	×	×	×	×
El Salvador	×	×	×	×	×	×
Mozambique	×	×	×	×	×	×
Niger	×	×	×	×	×	×
oPt	×	×	×	×	×	×
Philippines	×	×	×	×	×	×
Sudan	×	×	×			×

Qualitative

The qualitative data were received from five countries: Cambodia, El Salvador, Mozambique, Philippines, and Sudan. The data were analyzed through a systematic manual process, with pre-determined themes of feasibility and appropriateness guiding the analysis. Initially, responses were organized by question, grouping the answers from each office together to maintain a coherent structure for analysis. The data were then carefully reviewed, with responses manually sorted into the pre-determined categories of feasibility and appropriateness, along with their corresponding sub-themes. This was achieved by highlighting key points. A summary of the findings was then compiled, focusing on the frequency and significance of the identified sub-themes across different contexts. This process allowed for a comprehensive understanding of user perceptions.

A systematic approach was applied to ensure a thorough analysis of the data provided by enumerators through Kobo during the data collection phase. The comments were first compiled, translated, and cleaned to maintain accuracy. These comments were then disaggregated by country and coded according to the identified themes related to feasibility and appropriateness. The coded responses were analyzed using pivot tables, which allowed for the organization and comparison of data across different countries. The analysis particularly focused on identifying the top two challenges observed by enumerators in each country, providing a clear understanding of the most significant barriers to feasibility and appropriateness in various contexts. This method ensured that the key issues were highlighted and could be addressed in future tool iterations.

LIMITATIONS

There are several limitations of this study that are worth noting. In Mozambique, due to low performance of some enumerators, half of the enumerators were replaced between test-retest observations. As such, the low test-retest correlations in Mozambique may be partly due to enumerator differences, not time differences. In Sudan, data collection was paused due to the civil war and therefore only inter-rater reliability and CFA data could be collected. There also was the possibility of greater than average parental involvement in the assessment process. For example, in Cambodia enumerators shared that although the project team clearly articulated the purpose of the ReAL validation, some caregivers attempted to whisper answers and correct the children's mistakes during data collection and instances of children being blamed were also overheard during the on-call assessments. However, most countries reported none to minimal parental involvement (i.e., in Niger enumerators estimated approximately 5 percent of parents attempted to influence the results). We cannot confirm the level of involvement and propose measuring this in future studies.

ETHICS & ACCOUNTABILITY

This study received institutional review board (IRB) approval from Save the Children's Ethics Review Committee under protocol: SCUS-ERC-FY2023-9.

Enumerator Training

Enumerators were trained in a multi-day training by in-country trainers who had received a training-of-trainers curriculum delivered by the ReAL Global Team. The training workshop focused on familiarizing the enumerators with the child safeguarding protocol, tool modality, interviewing techniques, scoring guidelines, and data collection using a form programmed in KoBo Toolbox (KoBo Toolbox, n.d.).

Data Handling

Data were collected remotely via phone calls, text messages, and Save the Children staff and hired enumerators entered the data using the KoBo Toolbox software. The KoBo account was password-protected and managed by one of the principal investigators. After data collection, phones did not display the data collected but only the number of assessments carried out.

Access to the data in the server was available to one of the principal investigators and the in-country monitoring, evaluation, accountability, and learning (MEAL) team/academic partner staff. The MEAL team/academic partner staff monitored the data and ran quality checks with regular check-ins with the ReAL Global Team for troubleshooting. After all data were uploaded to the server, MEAL staff downloaded the data into a password-protected computer and separated names from the rest of the dataset. The de-identified dataset was shared with the other global co-PIs on the project via an encrypted file on Sharepoint.

Consent and Confidentiality

We used a process of verbal informed consent for the primary caregiver for her/his own and the child's participation, as well as verbal assent from child participants. Consent forms were written to explain the information with accuracy and clarity for individuals with low levels of literacy (2 years of formal education). Verbal consent was taken from the primary caregiver, and verbal assent was taken from children. All subjects were free to withdraw from the study at any time and were assured this would not affect the standard of care received in the community. During the consent process, consent forms were read aloud to the participants by data collectors. Data collectors were trained to have a conversation with the family to ensure each component had been understood and to discuss any questions that might arise. Each pilot site contextualized and translated accordingly.

Tom Maguire/Save the Children

FINDINGS

SUMMARY OF FINDINGS

The primary goal of the current study was to assess the psychometric properties of the High Access and Caregiver Report modalities of ReAL. Specifically, we appraised the inter-rater reliability, underlying factor structure, criterion validity, test-retest reliability, and item slope and difficulty. Our results show, with only a few notable exceptions, moderate to strong evidence that ReAL is a valid and reliable measure of the literacy and numeracy sub-domains assessed. The evidence for the social-emotional sub-domains is less robust or lacking.

A secondary objective of this study was to explore the perceived feasibility and appropriateness of using ReAL in different operating contexts. Specifically, we evaluated perceptions of feasibility by examining factors such as technical infrastructure, logistical challenges, the assessment environment, staff training, and scalability. We also assessed perceptions of appropriateness through considerations of contextualization, content relevance, and suitability. The qualitative findings revealed that while the tool is generally perceived as scalable and contextually appropriate, challenges persist with unreliable connectivity, caregiver influence, and comprehension issues in rural and linguistically diverse settings. Additionally, concerns were raised about the tool's relevance to the assessed grade levels indicating a need for further contextual adaptation.

Table 3
ReAL Interrater Reliability by Domain and Country

ReAL Domain	Percent Agreement (%)						
	Cambodia	El Salvador	Mozambique	Niger	oPt	Philippines	Sudan
	High Access (n=208)	Caregiver Report (n=208)	High Access (n=79)	High Access (n=219)	High Access (n=185)	High Access (n=92)	High Access (n=355)
Literacy	98.34	97.47	93.07	97.62	98.92	97.04	93.15
Numeracy	97.4	98.35	94.03	94.46	99.06	96.42	88.4
SEL	93.17	94.17	90.62	94.17	97.27	92.09	86.41

RQ1: WHAT ARE THE PSYCHOMETRIC PROPERTIES OF REAL TO REMOTELY ASSESS FOUNDATION SKILL DEVELOPMENT IN SEVEN LMICS?

In this section we present the results for each statistical analysis providing country-specific results under each analysis type. We present the statistics along with a qualitative interpretation of those statistics.

Interrater Reliability (IRR)

In the following section, we present the IRR results by country. We use classification guidelines as follows (Graham et al., 2012):

90% or above = High

75–89% = Acceptable

<75% = Unacceptable

For this study, IRR is a measure of the consistency of agreement between two assessors of the same child on a subsample within each country. IRR is critical for ensuring a measure is valid; a measure with low IRR would mean that there is a risk of assessors misclassifying a response.

IRR was high across nearly all domains, with the exception of the numeracy and SEL domains in Sudan, which had acceptable percent agreement. We present the domain-level IRR results in Table 3 below. Sub-domain-level results are in Appendix B.

Confirmatory Factor Analysis (CFA)

In the following section, we present CFA results by sub-domain of literacy, numeracy, and SEL, across countries. Country-specific fit statistics are found in Appendix C. In appraising the fit statistics, we use the following cutoff guidelines (also noted in the quantitative analysis section):

- Values greater than 0.90 for the CFI are considered to reflect acceptable and excellent fit to the data (Marsh et al., 2005)
- Values for RMSEA less than 0.05 and 0.08 indicate excellent fit and acceptable fit to the data, respectively (Hu & Bentler, 1999)
- A SRMR value less than .08 (Hu & Bentler, 1999)

CFA is an important analysis to conduct, as it provides us with an understanding of whether the construct, or skill, that we hypothesized would be measured by the items for a given sub-domain actually measure the construct. We look at the fit statistics presented in Table 4 and use the cutoff guidelines above to apprise goodness of fit.

We did not conduct CFA on expressive vocabulary, retelling a story, one-to-one correspondence, or word problems due to too few items for CFA. These CFA results do not include El Salvador.

As presented in Table 4, we find satisfactory model fit for all sub-domains in literacy, except sentence-level comprehension. All sub-domain CFA models in numeracy exhibited satisfactory fit indices. The results for the SEL sub-domains are mixed: fit statistics for use of social supports and help seeking behavior were only partially satisfactory and unsatisfactory for all others.

Table 4
CFA Fit Statistics for Each Sub-domain of Literacy, Numeracy, and Social and Emotional Learning

Sub-domain	$\chi 2$(df)	*p*	RMSEA [90%CI]	CFI	SRMR	Fit Appraisal
Literacy						
Listening comprehension	5.966 (5)	.310	.033 [.000, .090]	.999	.007	Satisfactory
Letter/letter sound identification	652.951 (170)	< .001	.155 [.143, .167]	.886	.033	Partially satisfactory
Common word identification	309.506 (35)	< .001	.158 [.141, .176]	.946	.023	Satisfactory
Sentence–level comprehension	743.763 (5)	< .001	.754 [.690, .820]	.436	.184	Unsatisfactory
Reading comprehension	188.439 (14)	< .001	.181 [.156, .207]	.862	.066	Partially satisfactory
Numeracy						
Number identification	468.752 (54)	< .001	.040 [.037, .044]	.994	.049	Satisfactory
Addition	283.185 (35)	< .001	.140 [.123, .158]	.951	.026	Satisfactory
Subtraction	353.667 (35)	< .001	.175 [.156, .195]	.921	.035	Satisfactory
Social and emotional learning						
Self-concept[1]	Heywood					Unsatisfactory
Use of social supports[2]	0.935 (2)	.627	.000 [.000, .000]	1.000	.152	Partially satisfactory
Help seeking behavior[3]	0.612 (2)	.736	.000 [.000, .020]	1.000	.142	Partially satisfactory
Stress management[4]	Heywood					Unsatisfactory
Empathy[5]	2443.116 (43)	< .001	.108 [.105, .112]	.971	.153	Un satisfactory
Conflict resolution[6]	Heywood					Unsatisfactory

Note
Robust RMSEA and robust CFI were reported for listening comprehension, letter/letter sound identification, common word identification, sentence–level comprehension, oral passage reading, addition, subtraction, and use of social supports according to Savalei (2021).

1
Six items (sel1–sel6) loaded on the latent factor. The tetrachoric correlations between items ranged from .94 to .99. sel6 exhibited negative variance.

2
Four items (rel3, rel4, rel9, and rel13) loaded on the latent factor. The tetrachoric correlation coefficient between rel3 and rel4 was .99.

3
Four items (rel5, rel6, rel10, and rel14) loaded on the latent factor. The tetrachoric correlation coefficient between rel5 and rel6 was .99.

4
Four items (rel1, rel2, rel8, and rel12) loaded social support, and three items loaded behavioral regulation (st1–st3). The tetrachoric correlation coefficient between rel1 and rel2 was .99. The tetrachoric correlation coefficients between st1, st2, and st3 ranged from .98 to .99. st3 exhibited negative variance.

5
Five items (rel7, rel11, rel15, e1, and e6) loaded on identifying feelings of others, and six items (e2, e3, e5, e7, e8, and e10) loaded on empathy.

6
Four items (con1–con4) loaded on social problem solving, and two items (e4 and e9) loaded on interpreting hostility. The tetrachoric correlation coefficients between con1–con4 ranged from .93–.98. con4 exhibited negative variance.

Table 5 shows the CFA fit statistics for El Salvador, where the ReAL Caregiver Report modality was tested. We find satisfactory or partially satisfactory fits for most sub-domains in the SEL domain, with the exception of conflict resolution. The model converged with partially satisfactory fit statistics for literacy. However, the model did not converge for the numeracy domain.

Table 5
CFA Fit Statistics for Each Domain of Literacy, Numeracy, and Social and Emotional Learning for El Salvador

Sub-domain	χ2(df)	*p*	RMSEA [90%CI]	CFI	SRMR	Fit Appraisal
Literacy[1]	9747.87 (45)	<.001	.063 [.053, .073]	.988	.064	Partially satisfactory
Numeracy[2]	Heywood					Unsatisfactory
Social and emotional learning						
Self-concept[3]	11.625 (5)	.040	.036 [.007, .064]	1.000	.019	Satisfactory
Use of social supports[4]	3.557 (2)	.169	.028 [.000, .074]	.972	.285	Partially satisfactory
Help seeking behavior[5]	3.471 (2)	.176	.027 [.000, .073]	.977	.292	Partially satisfactory
Stress management[6]	1.084 (2)	.581	.000 [.000, .052]	1.000	.275	Partially satisfactory
Empathy[7]	115.134 (26)	<0.001	.252 [.205, .300]	.739	.079	Partially satisfactory
Conflict resolution[8]	Heywood					Unsatisfactory

1
A correlated four-factor CFA model was developed for literacy: ev1 and ev2 loaded on expressive vocabulary; rs1 and rs2 loaded on retelling story; l11, l12, and l13 loaded on letter/letter sound identification; w1, w2, and w11 loaded on common word. The seven latent factors correlated.

2
A correlated three-factor CFA model was developed for numeracy: o1, o2, and o3 loaded on one-to-one correspondence; n1, n2, n3, and n4 loaded on number identification; add1 and add2 loaded on addition. The three latent factors correlated. This model produced negative variance for o3. After removing o3 from the model, o2 exhibited negative variance. Thus, a correlated two-factor CFA model was developed: n1, n2, n3, and n4 loaded on number identification; add1 and add2 loaded on addition. However, the correlated two-factor CFA model had convergence problem.

3
A one-factor CFA model was developed for self-concept: sel1–sel6 loaded on one latent factor. This model produced negative variance for sel1. A one-factor CFA model with sel2–sel6 was developed.

4
A one-factor CFA model was developed for use of social supports: rel3, rel4, rel9, and rel13 loaded on one latent factor

5
A one-factor CFA model was developed for help seeking behavior: rel5, rel6, rel10, and rel14 loaded on one latent factor.

6
A correlated two-factor CFA model was developed for stress management: rel1, rel2, rel8, and rel12 loaded on social support; st1, st2, and st3 loaded on behavioral regulation. The two latent factors correlated. This model produced negative variance for st1 and st2.

7
A correlated two-factor CFA model was developed for stress management: rel7, rel11, rel15, e1, and e6 loaded on identifying feelings of others; e2, e3, e5, e7, e8, and e10 loaded on empathy. The two latent factors correlated. This model produced negative variance for e1. The correlated two-factor model fitted to the data after removing e1. This model produced negative variance for e6. Thus, e6 was removed from the model. The correlated two-factor model became: rel7, rel11, and rel15 loaded on identifying feelings of others; e2, e3, e5, e7, e8, and e10 loaded on empathy. The correlation coefficient between the two factors was .402.

8
A correlated two-factor CFA model was developed for stress management: con1, con2, con3, and con4 loaded on social problem solving; e4 and e9 loaded on interpreting hostility. The two latent factors correlated. This model produced negative variance for con1.

Criterion Validity

In the following section, we present the criterion validity results by sub-domain and country, with the exclusion of El Salvador for which we did not have a valid criterion measure. In assessing the Spearman correlation coefficients, we use the following guidelines:

Satisfactory (> .2)

Fair (.1 – .19)

Unsatisfactory (<.1)

Inconclusive (mixed results, close to .1)

We measured criterion validity to understand whether ReAL relates to other validated, established measures of the same skills. In Table 6 below, we present the Spearman correlation coefficients along with color-coded assessments of the strength of the association between ReAL and the other assessments used.

Country-specific results are shown, as the criterion measures in each country differed. For sub-domains in which there was no associated construct being measured on the criterion measure, we were unable to run a correlation.

In the literacy domain, we find strong positive and significant correlations for the majority of tested sub-domains in Cambodia, Niger, and oPt. For Mozambique, the results are more mixed, while in the Philippines there is not a strong relationship on any of the sub-domains. In terms of the numeracy criterion correlations, a positive correlation was seen between all ReAL sub-domains and corresponding criterion sub-domains, except for one-to-one correspondence which was not included due to lack of variability and lack of a corresponding criterion measure. Criterion correlations for the SEL domain were inconclusive; correlations were often negative, in an unexpected direction, and were mostly insignificant.

Table 6
ReAL-Criterion Measure Spearman Correlation Coefficients & Classification by Sub-domain and Country

	Cambodia	Mozambique	Niger	oPt	Philippines
	High Access	High Access	High Access	High Access	High Access
Literacy	(n=206)	(n=79)	(n=217)	(n=65)	(n=188)
Expressive Vocabulary	Not tested	Not tested	Not tested	Not tested	Not tested
Retelling Story	Not tested	Not tested	Not tested	Not tested	Not tested
Listening Comprehension	Not tested	Not tested	Not tested	Not tested	Not tested
Letter/Letter Sound Identification	.70*, .69*, .72*	.37	.44*	–.04	–.06, –.10
Common Word Identification	.78*, .67*	.13	.46*	.53	–.08
Sentence Comprehension	Not tested	Not tested	Not tested	Not tested	Not tested
Reading Comprehension	.72*	.06	.26*	.50	.17*
Numeracy	(n=206)	(n=79)	(n=217)	(n=65)	(n=188)
One to One Correspondence	Not tested	Not tested	Not tested	Not tested	Not tested
Number Identification	.78*	Not tested	.34*	Not tested	Not tested
Addition	.74*	.20	.18	.25	.10, .18*
Subtraction	.70*	.20	.08	.39	.16*, .09
Word Problems	Not tested	.13	Not tested	.54	.21*
Social and Emotional Learning	(n=201)	(n=79)	(n=217)	(n=65)	(n=188)
Self-Concept	Not tested	0	Not tested	0	.02
Use of Social Supports	.08	–.09	.06	.02	–.05
Help Seeking Behavior	.13	–.06	.15*	–.04	0
Stress Management, Social Supports	.06	.27, –.01	–.03, .09	.19, .05	–.11
Stress Management, Behavioral Regulation	Not tested	–.05, .30	–.03, .05	.14, –.15	–.01
Identifying Feelings of Others	.04	.08, –.06	–.01, .12	.18, –.02	–.07, .04
Empathy	.02, .40*	–.10, .24	.05, .09	.22, .06	–.08
Conflict Resolution	Not tested	Not tested	Not tested	Not tested	Not tested

* indicates significance at alpha= .05,
only calculated for countries with sample sizes >100

Test-Retest Reliability (TRT)

In the following section, we present the test-retest reliability results by sub-domain and country. In assessing the Spearman correlation coefficients, we use the following guidelines (Cicchetti, 1994):

Excellent = .75 and above

Good = .60 – .74

Fair = .40 – .59

Low = .39 and below

In measuring test-reliability, we are seeking to understand whether the results from one administration of ReAL can be reproduced in the same population at a second time point. For an assessment like ReAL, this is very important, as we want to understand how consistent it is in measuring the same skill over time.

Table 7 shows the Spearman correlation coefficients for test-retest reliability. For the literacy domain, we find that correlations ranged from fair to excellent for most countries (Cambodia, Niger, and oPt) with only unsatisfactory correlations on two sub-domains across Niger and oPt.

For Mozambique, El Salvador, and the Philippines, reliability was more mixed, with the Philippines having the lowest test-retest reliability across sub-domains. For numeracy, Cambodia, El Salvador, and oPt had the highest overall reliability ranging from fair to excellent (with the exception of one-to-one correspondence in El Salvador and Cambodia, which had low reliability). Mozambique, Niger, and the Philippines had low to good reliability across most sub-domains. For SEL, Cambodia, El Salvador, and oPt had mixed results ranging from fair to good reliability for most sub-domains and low for the rest. Reliability in Mozambique, Niger, and the Philippines was either completely or mostly low across sub-domains.

Table 7
ReAL Spearman Correlation Coefficients for Test-Retest Reliability by Sub-domain and Country

	Cambodia		El Salvador		Mozambique		Niger		oPt		Philippines	
	High Access		Caregiver Report		High Access		High Access		High Access		High Access	
Literacy	Sample	Corr.	Sample	Corr.	Sample	Corr.	Sample	Corr.	Sample	Corr.	Sample	Corr.
Expressive Vocabulary	210	.64	141	.15	77	.09	171	.48	168	.65	153	.40
Retelling Story	159	.46	135	.44	6	.79	82	.52	146	.38	92	.33
Listening Comprehension	210	.49	138	.08	54	.23	154	.44	163	.54	153	.14
Letter Identification	210	.74	141	.68	78	.47	188	.42	168	.77	153	–.02
Common Word Identification	198	.83	130	.81	42	.60	184	.51	156	.70	153	.33
Sentence Comprehension	189	.86	140	.66	75	.28	184	.33	158	.53	153	.48
Reading Comprehension	126	.60	140	.80	14	.09	85	.52	106	.55	138	.48
Numeracy	Sample	Corr.	Sample	Corr.	Sample	Corr.	Sample	Corr.	Sample	Corr.	Sample	Corr.
One to One Correspondence	210	.30	141	.37	79	N/A	182	.01	167	.41	153	–.01
Number Identification	209	.80	140	.83	80	.68	186	.49	166	.79	152	–.05
Addition	194	.78			70	.46	185	.33	155	.41	152	.20
Subtraction	196	.76	141	.81[1]	63	.34	185	.28	156	.70	152	.15
Word Problems	192	.66	138	.75	68	.30	172	.19	156	.58	152	.38
Social and Emotional Learning	Sample	Corr.	Sample	Corr.	Sample	Corr.	Sample	Corr.	Sample	Corr.	Sample	Corr.
Self-Concept	210	.46	141	.49	80	.29	192	.25	169	–.03	153	.56
Use of Social Supports	210	.56	141	.39	80	–.13	192	.09	169	.36	153	.28
Help Seeking Behavior	210	.65	141	.40	80	–.04	192	.13	169	.55	153	.28
Stress Management, Social Supports	210	.53	141	.46	80	.13	192	.07	169	.43	153	.46
Stress Management, Behavioral Regulation	210	.37	141	.13	80	–.04	192	.13	169	.16	153	.05
Identifying Feelings of Others	207	.50	141	.47	79	–.15	174	.22	165	.46	151	.29
Empathy	206	.62	141	–.38	79	–.01	174	–.13	165	.60	137	.14

1
Addition and subtraction were combined in the Caregiver Report

Item and Test Characteristics

In the following section, we discuss the results of the IRT analysis, referencing the IIFs and TIFs found in Appendices E and F. We conducted IRT analyses and plotted the IIFs and TIFs in order to understand how ReAL differentiates between the skill levels of those children assessed and for which children the assessment provides the most reliable information.

Despite the variability of items in distinguishing between children with different skills, the peaks for all IIFs were located at least one standard deviation below the mean. This suggests that these sub-domains were easy for the participants. Consequently, each sub-domain merely provides reliable information for children whose skills are lower than average. This is evidenced by the TIFs. As a result, all sub-domains of literacy and numeracy may not be able to accurately assess skills for the majority of children.

RQ2: WHAT ARE THE PERCEPTIONS OF USERS OF THE ReAL TOOL DURING THE VALIDATION STUDY ABOUT ITS FEASIBILITY AND APPROPRIATENESS IN DIFFERENT OPERATING CONTEXTS?

In this section, we present the results of the qualitative analysis and illustrate our findings with relevant examples.

Perceptions Around Feasibility

Users of the ReAL tool across various operating contexts shared their perceptions regarding its feasibility. A consistent finding was that access to devices was not a significant issue; however, unreliable internet and phone networks presented varying degrees of challenges. For example, respondents from Philippines mentioned that some areas were beyond the reach of cellular signals.

Additionally, respondents from different contexts noted that the influence of caregivers during the assessment process impacted the responses. In El Salvador, maintaining the focus of caregivers for an extended period of time proved challenging, while in Cambodia, instances of fear, anxiety, and concerns about caregivers potentially criticizing children's learning abilities were observed. This issue was also identified as one of the top challenges by enumerators from Cambodia, the Philippines, oPt, and Mozambique. Enumerators from oPt noted that caregivers, particularly parents or siblings, were involved in helping children answer the assessment questions, which may have influenced the accuracy of the responses.

Despite these challenges, all respondents expressed a belief in the high scalability potential of the tool within their countries. However, concerns were raised regarding the tool's ability to effectively reach vulnerable populations, particularly in rural areas with limited connectivity.

Perceptions Around Appropriateness

Regarding the ReAL tool's appropriateness, respondents from five countries, with only a few notable exceptions, agreed that the questions in the tool were effectively contextualized and resonated well with the target populations. One exception being, enumerators in oPt reporting that many children struggled to understand questions in the emotion identification section of the SEL domain, which led to an increase in clarifying questions and subsequently extended the interview time. Additionally, in El Salvador and Mozambique, some difficulties in comprehension were encountered, particularly in rural settings and communities with linguistic variations. This issue was also highlighted by enumerators in Mozambique and Niger, where assessments were conducted in Portuguese and French—languages not fully understood by the respondents and caregivers. Furthermore, respondents from Sudan, Philippines, and Cambodia raised concerns that some of the items were no longer relevant for the grade level being assessed, which could impact the accuracy of understanding the learning levels based on the assessment.

Tom Maguire / Save the Children

DISCUSSION, IMPLICATIONS, & FUTURE DIRECTIONS

The goal of the current study was to assess the psychometric properties of the High Access modality of ReAL and to gather some initial evidence for the Caregiver Report version. Specifically, we appraised the inter-rater reliability, underlying factor structure, item slope and difficulty, criterion validity, and test-retest reliability. Our results show, with only a few notable exceptions, moderate evidence that the ReAL High Access version is a valid and reliable measure of the literacy and numeracy sub-domains assessed. The evidence for the social-emotional sub-domains is less robust or lacking. Similarly, users of the ReAL tool also recommend further contextualization of questions to better suit various grade levels. From these results, we propose revisions to the literacy and numeracy items to bring them more in line with the skills levels of 5–14-year-old children in the countries that participated in this study. We also recommend redeveloping the SEL sub-domains and piloting again.

Implications of Different Results by Academic Versus Non-academic Domains

The results for literacy and numeracy demonstrate strong evidence for valid and reliable measures of these skills. With the exception of expressive vocabulary, retelling a story, sentence-level comprehension, and one-to-one correspondence, we find reasonable inter-rater reliability, construct validity, criterion validity, and test-retest reliability. The IIFs and TIFs suggest that making items more difficult in these sub-domains while retaining the assessment approach could precisely assess a wide range of ability.

The results for social and emotional skills are less optimistic. Despite many sub-domains performing well in face-to-face assessments (D'Sa & Krupar, 2021; Krupar & D'Sa, 2024), there is inconclusive or poor evidence supporting the current form of social and emotional sub-domains in the ReAL tool. Numeracy and literacy skills are easier to assess than social and emotional skills due to their straightforward and standardized nature, allowing for consistent measurement and comparability across different countries and cultures (Kaffenberger & Pritchett, 2020).

Social and emotional skills are context-dependent, requiring significant adaptation to be accurately assessed in different cultural settings (Schonert-Reichl et al., 2015). The subjective nature of these assessments introduces additional complexity, emphasizing the need for comprehensive enumerator training (Humphrey, 2013). Enumerators in Sudan shared that although the questions asked in the social and emotional skills domain were child-friendly, they were quite sensitive for children who had experienced conflict in their lifetime. Fear or a desire to appear "normal" might have led the children to downplay their true feelings and experiences, especially to enumerators that were unfamiliar to the children. In Niger, enumerators shared that social and emotional skills are not integrated into the education curriculum, children are not accustomed to being questioned about their social-emotional wellbeing, and, as a result, many children remained silent during questioning for this section, likely due to lack of understanding or not knowing how to respond. Enumerators shared a similar sentiment in Cambodia and spoke to how social-emotional learning is only integrated into the curriculum in specific UNICEF-funded schools, and most schools do not teach any social-emotional learning curriculum. Colleagues in the Philippines shared that the SEL curriculum has not been mainstreamed into the education framework, that teachers are not trained in SEL and there are uncoordinated efforts among sectors and agencies. Given that the ReAL tool is designed to access hard- to- reach children, these considerations will be central for future iterations and revisions of the tool.

Administering these assessments remotely further complicates the process, especially for SEL. Remote assessments lack the direct, in-person interaction crucial for accurately gauging social and emotional competencies. Enumerators may struggle to interpret non-verbal cues or create a supportive environment conducive to honest self-expression when not physically present. This challenge is compounded by the lack of substantial evidence on the effectiveness and reliability of remote SEL assessments, as most existing studies and tools are designed for in-person administration (Domitrovich et al., 2017). This shift highlights the urgent need for research to develop and validate reliable methods for remotely assessing SEL skills. However, the complications due to the remote administration influence all domains of ReAL. For example, enumerators in Mozambique shared that although children often recognize lowercase and uppercase letters when they are hand written, they may not be able to recognize the typed letter on a screen.

Implications of Limited Variability Across Target Age Range

Appraising the IIFs and TIFs in our study shows that several items and sub-domains mainly provide information for children performing below the mean, particularly younger ones. This phenomenon is most apparent among younger children due to their varying stages of cognitive and emotional development, which influence their learning abilities and skill acquisition. Younger children have not yet progressed through higher order developmental stages and may advance through them at different times, which can lead to significant differences in their literacy, numeracy, and social and emotional skills. Furthermore, the early years are crucial for language development, and children's exposure to language-rich environments significantly impacts their literacy skills (Shonkoff & Phillips, 2000). Similarly, early experiences with numbers and spatial reasoning, which vary widely depending on the child's home environment and parental engagement, affect numeracy skills (Ginsburg et al., 2008).

Social and emotional skills are context-dependent and develop through interactions with caregivers and peers, leading to variability in children's social competencies and emotional regulation skills (Denham et al., 2012). Younger children are more susceptible to adverse experiences like poverty and stress, impacting their development in these domains (Yoshikawa et al., 2012). As children grow older, more structured education helps mitigate early disparities. Older children have had more time to develop compensatory skills and strategies, reducing variability and disparities seen in younger age groups (Entwisle & Alexander, 1993). In contrast, foundational skills in literacy and numeracy begin with lower-order skills essential for acquiring higher-order competencies. In literacy, this starts with phonemic awareness and phonics (National Institute of Child Health and Human Development, 2000). As children progress, they develop higher-order literacy skills, such as fluency, vocabulary, and comprehension, enabling them to read with speed, accuracy, and expression, and to understand and interpret complex texts (Snow, 2002). In numeracy, lower order skills begin with number sense and basic arithmetic (Berch, 2005). We propose retaining the wide age range to increase applicability in LMIC contexts, as older children may still be mastering foundational skills due to lack of access or disruptions to school. We also suggest revising the measure to include more difficult items for higher-performing children while dropping the sub-domains retelling a story, sentence-level comprehension, and one-to-one correspondence due to their poor performance across age groups. We also propose exploring the feasibility of making ReAL an adaptive assessment.

Implications and Future Directions

Through this study, we demonstrate that a remote assessment of learning can be a feasible, valid, and reliable measure of foundational academic (i.e., literacy and numeracy) skills in LMICs. While this evidence is promising, there are critical preconditions that must be met when conducting a remote, phone-based assessment: A cellular network infrastructure and connectivity and ownership of or access to a phone. These conditions may not always be present in many low-resource contexts. For this reason, it is important to understand the context in which the assessment will take place to evaluate whether such an assessment is appropriate. This will also aid in identifying the linguistic background of the participants, allowing for the tools to be accurately translated and adapted to the specific language needs of the context. Thus, we advocate for assessments like ReAL to be added to the other options educations systems have to assess foundational skills rather than serving as a replacement. We also identified one other precondition for the successful implementation of the tool: constructive caregiver support. Addressing the role of caregivers during assessments may involve creating guidelines or training materials to ensure that this involvement is guided and managed appropriately, enhancing rather than detracting from the accuracy and reliability of the data collected.

Given the promise of remote, phone-based assessments as one option to assess foundational skills of hard-to-reach children in LMIC contexts, we advocate for further research that builds upon this validation study. Specifically, we propose developing and testing adaptive versions that could more efficiently assess skills across the wide age range we target here and piloting the tool in more diverse settings to refine its reach and effectiveness, particularly among vulnerable populations in rural areas. Additionally, building evidence of cost effectiveness in comparison with other assessment types will be a critical consideration for any education system weighing a phone-based assessment like ReAL with one administered face-to-face. Finally, we seek to make the literacy and numeracy items more difficult and to re-develop and test the ReAL SEL sub-domains in future studies.

REFERENCES

Aker, J. C., & Ksoll, C. (2020). Can ABC lead to sustained 123? The medium-term effects of a technology-enhanced adult education program. *Economic Development and Cultural Change*, 68(3), 1081-1102.

Angrist, N., Bergman, P., Evans, D. K.,Hares, S., Jukes, M. C. H., & Letsomo, T. (2020). Practical lessons for phone-based assessments of learning. *BMJ Global Health*, 5, e003030.

Barbot, B., Hein, S., Trentacosta, C., Beckmann, J. F., Bick, J., Crocetti, E., … van IJzendoorn, M. H. (2020). Manifesto for new directions in developmental science. *New Directions for Child and Adolescent Development*, 135–149. https://doi.org/10.1002/cad.20359

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

Berch, D. B. (2005). Making sense of number sense: Implications for children with mathematical disabilities. *Journal of Learning Disabilities*, 38(4), 333-339.

Collaborative for Academic, Social, and Emotional Learning. (n.d.). What is the CASEL framework? Fundamentals of SEL. CASEL. Retrieved July 9, 2024, from https://casel.org/fundamentals-of-sel/what-is-the-casel-framework/

Chalmers, R. P. (2012). mirt: A multidimensional item response theory package for the R environment. *Journal of Statistical Software*, 48(6), 1–29. https://doi.org/10.18637/jss.v048.i06

Denham, S. A., Bassett, H. H., & Wyatt, T. (2012). The socialization of emotional competence. In *J. Grusec & P. Hastings (Eds.), Handbook of Socialization: Theory and Research* (2nd ed., pp. 614-637). Guilford Press.

Domitrovich, C. E., Durlak, J. A., Staley, K. C., & Weissberg, R. P. (2017). Social-emotional competence: An essential factor for promoting positive adjustment and reducing risk in school children. Child Development, 88(2), 408–416. https://doi.org/10.1111/cdev.12739

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

D'Sa, N., Krupar, A., & Westrope, C. (2019). Feasible measurement of learning in emergencies: lessons from Uganda. *Forced Migration Review: Education needs, rights and action in displacement* (Issue March). ISELA : https://inee.org/resources/international-social-and-emotional-learning-assessment-isela

D'Sa, N., & Krupar, A. (2021). Developing and Validating the International Social and Emotional Learning Assessment: Evidence from a Pilot Test with Syrian Refugee Children in Iraq. *Journal on Education in Emergencies*, 7(2): 20-56. https://doi.org/10.33682/xdpq-bwp2.

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.

Entwisle, D. R., & Alexander, K. L. (1993). Entry into school: The beginning school transition and educational stratification in the United States. *Annual Review of Sociology*, 19, 401-423.

Ginsburg, H. P., Lee, J. S., & Boyd, J. S. (2008). Mathematics education for young children: What it is and how to promote it. *Social Policy Report*, 22(1), 3-22.

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., Wolf, S., Yoshikawa, H., Rojas, N., Kabay, S., Dowd, A. J., & Pisani, L. (2019). Measuring early learning and development across cultures: Invariance of the IDELA across five countries. *Developmental Psychology*, 55(1), 23-37.

Hu, L. T., & 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

Humphrey, N. (2013). Social and emotional learning: A critical appraisal. *SAGE Publications Ltd.*

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.

Kaffenberger, M., & Pritchett, L. (2020). Aiming higher: Learning outcomes in low- and middle-income countries. International Journal of Educational Development, 78, 102242. https://doi.org/10.1016/j.ijedudev.2020.102242

KoBo Toolbox. (n.d.). Kobo Toolbox | Data Collection Tools for Challenging Environments. Kobo Toolbox. Retrieved July 24, 2024, from https://kobotoolbox.org/.

Krupar, A., D'Sa, N., Westrope, C., & Johna, J. F. (2019). Developing a Holistic Assessment of Research-Practice Partnership Narrative Westrope, Ponguta, Hein, Schubert Children's Learning in the Context of Forced Displacement: Case Study from Dadaab, Kenya. In Data Collection and Evidence Building to Support Education in Emergencies. NORRAG, 2, 63–65.

Krupar, A., & D'Sa, N. (2024). Measuring learning during crises: Developing and validating the Holistic Assessment of Learning and Development Outcomes (HALDO). *International Journal of Educational Research Open*, 6(2024), 100320. https://doi.org/10.1016/j.ijedro.2024.100320

Marsh, H. W., Hau, K.-T., & Grayson, D. (2005). Goodness of fit evaluation in structural equation modelling. In A. Maydeu-Olivares & J. J. McArdle (Eds.), Contemporary psychometrics: A festschrift for Roderick P. McDonald (pp. 275–340). Erlbaum.

Montoya, S., Mundy, K., and Scheid, P. (2016). Understanding what works in oral reading assessments. Global Partnership for Education. Retrieved from https://www.globalpartnership.org/blog/understanding-what-works-oral-reading-assessments

Mulligan, C. A., & Ayoub, J. L. (2023). Remote assessment: Origins, benefits, and concerns. *Journal of Intelligence*, 11(6), 114. https://doi.org/10.3390/jintelligence11060114

National Institute of Child Health and Human Development (2000). Report of the National Reading Panel. Teaching children to read: An evidence-based assessment of the scientific research literature on reading and its implications for reading instruction.

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

Olayemi, M., Tucker, M., Choul, M., Purekal, T., Benitez, A., Wheaton, W., & DeBoer, J. (2021). Creating a Tool to Measure Children's Wellbeing: A PSS Intervention in South Sudan. *Journal on Education in Emergencies* 7(2): 104-51. https://doi.org/10.33682/rhqb-fy8u.

Panter-Brick, C., Hadfield, K., Dajani, R., Eggerman, M., Ager, A., & Ungar, M. (2018). Resilience in context: A brief and culturally grounded measure for Syrian refugee and Jordanian host-community adolescents. *Child Development*, 89(5), 1803-1820. https://doi.org/10.1111/cdev.12868

Pisani, L., Borisova, I., & Dowd, A. J. (2018). Developing and validating the International Development and Early Learning Assessment (IDELA). *International Journal of Educational Research*, 91, 1–15. https://doi.org/10.1016/J.IJER.2018.06.007

R Core Team (2021). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. https://www.R-project.org/.

Rapp, S. R., Legault, C., Espeland, M. A., Resnick, S. M., Hogan, P. E., Coker, L. H., Dailey, M., Shumaker, S. A., & CAT Study Group (2012). Validation of a cognitive assessment battery administered over the telephone. *Journal of the American Geriatrics Society*, 60(9), 1616–1623. https://doi.org/10.1111/j.1532-5415.2012.04111.x

Rosseel, Y. (2012). lavaan: An R package for structural equation modeling. *Journal of Statistical Software*, 48(2), 1–36. https://doi.org/10.18637/jss.v048.i02

RTI International. (2014). Early Grade Mathematics Assessments (EGMA) Toolkit. First Edition. Research Triangle Park, NC. RTI International. EGRA_Toolkit_Second_Edition_March_8_2016_Final_Internal_Edits23May2016_clean_HiRes_0.pdf (edu-links.org)

RTI International. (2015). Early Grade Reading Assessment (EGRA) Toolkit, Second Edition. Washington, DC: United States Agency for International Development. EGRA_Toolkit_Second_Edition_March_8_2016_Final_Internal_Edits23May2016_clean_HiRes_0.pdf (edu-links.org)

Sachs, J. D. (2012). From millennium development goals to sustainable development goals. *Lancet (London, England)*, 379(9832), 2206–2211. https://doi.org/10.1016/S0140-6736(12)60685-0

Savalei, V. (2021). Improving fit indices in structural equation modeling with categorical data. Multivariate Behavioral Research, 56(3), 390–407. https://doi.org/10.1080/00273171.2020.17179

Save the Children. (n.d.). *Remote assessment of learning: A real toolkit*. Save the Children Resource Centre. Retrieved July 9, 2024, from https://resourcecentre.savethechildren.net/document/remote-assessment-of-learning-real-toolkit/

Save the Children. (n.d.). Save the Children's literacy boost toolkit: Introduction. Save the Children. https://resourcecentre.savethechildren.net/document/save-childrens-literacy-boost-toolkit-introduction/

Save the Children. (n.d.). *Save the Children's numeracy boost toolkit: General overview*. Save the Children. https://resourcecentre.savethechildren.net/document/numeracy-boost-toolkit-general-overview/

Schonert-Reichl, K. A., Kitil, M. J., & Hanson-Peterson, J. (2015). To reach the students, teach the teachers: A national scan of teacher preparation and social and emotional learning. A report prepared for the Collaborative for Academic, Social, and Emotional Learning (CASEL). University of British Columbia. Retrieved from https://casel.org/wp-content/uploads/2016/01/SEL-TEd-Executive-Summary-for-CASEL-2015.pdf

Shavitt, I., Ayres de Araujo Scatollin, M., Suzart Ungaretti Rossi, A., Pacífico Mercadante, M., Gamez, L., Resegue, R.M., Pisani, L., Conceição do Rosário, M. (2022). Transcultural adaptation and psychometric properties of the International Development and Early Learning Assessment (IDELA) in Brazilian pre-school children. *International Journal of Educational Research Open*, *3*, 2666-3740. https://doi.org/10.1016/j.ijedro.2022.100138

Shonkoff, J. P., & Phillips, D. A. (2000). From Neurons to Neighborhoods: The Science of Early Childhood Development. National Academy Press.

Shumba, C.S. & Lusambili, A.M. Not enough traction: Barriers that aspiring researchers from low- and middle-income countries face in global health research. *Journal of Global Health Economics and Policy*. 2021;1:e2021002.doi:10.52872/001c.25802

Snow, C. E. (2002). Reading for Understanding: Toward an R&D Program in Reading Comprehension. RAND Corporation.

Sobers, S. M., Whitehead, H. L., N'Goh, K. N. A., Ball, M. C., Tanoh, F., Akpe, H., Jasinska, K. (2023). Is a phone-based language and literacy assessment a reliable and valid measure of children's reading skills in low-resource settings? *Reading Research Quarterly*, *58*(4), 733-754. https://doi.org/10.1002/rrq.511

Sowa, P., Jordan, R., Ralaingita, W., & Piper, B. (2021). Higher grounds: Practical guidelines for forging learning pathways in upper primary education. RTI Press. RTI Press Occasional Paper No. OP-0069-2105 https://doi.org/10.3768/rtipress.2021.op.0069.2105

Sticht, T. G, Hofstetter, C. R., & Hofstetter, C. H. (1996). Assessing adult literacy by telephone. *Journal of Literacy Research*, 28, 525–59.

Tubbs Dolan, C. (2019). *Psychometric analysis of the pilot Syria Holistic Assessment for Learning (SHAL): Validity and reliability*. New York University: Unpublished manuscript.

UNESCO. (2023). 250 million children out of school: What you need to know about UNESCO's latest education data. UNESCO. Retrieved July 9, 2024, from https://www.unesco.org/en/articles/250-million-children-out-school-what-you-need-know-about-unescos-latest-education-data

UNICEF. (2018). East Asia-Pacific Early Child Development Scales (EAP-ECDS). https://hkece.edu.hku.hk/content/uploads/2020/11/EAP-ECDS-Final-Report.pdf

UNICEF. (2021). COVID-19 and school closures: One year of education disruption. https://data.unicef.org/wp-content/uploads/2021/03/COVID19-and-school-closures-report.pdf

United Nations. (2023). Goal 4: Ensure inclusive and equitable quality education and promote lifelong learning opportunities for all. https://sdgs.un.org/goals/goal4

Wolf, S., Halpin, P., Yoshikawa, H., Dowd, A. J., & Pisani, L. (2017). Measuring school readiness globally: Assessing the construct validity and measurement invariance of the International Development and Early Learning Assessment (IDELA) in Ethiopia. *Early Childhood Research Quarterly, 41* (April), 21–36. https://doi.org/10.1016/j.ecresq.2017.05.001

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., Aber, J. L., & Beardslee, W. R. (2012). The effects of poverty on the mental, emotional, and behavioral health of children and youth: Implications for prevention. *American Psychologist*, 67(4), 272-284

Yoshikawa, H., Leyva, D., Snow, C. E., Treviño, E., Barata, M., Weiland, C., et al., (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.

APPENDICES

APPENDIX A: DESCRIPTIVE TABLES

Cambodia (n=1009)

Sex	Number	%
Male	480	47.57
Female	529	52.43
Missing	0	0.00
Age	**Number**	**%**
Average; SD	9.15	2.25
4	2	0.20
5	103	10.21
6	129	12.78
7	123	12.19
8	129	12.78
9	129	12.78
10	125	12.39
11	138	13.68
12	120	11.89
13	11	1.09
14	0	0.00
15+	0	0.00
Missing	0	0.00
School Year	**Number**	**%**
Kindergarten	124	12.29
1st grade	114	11.30
2nd grade	153	15.16
3rd grade	80	7.93
4th grade	144	14.27
5th grade	164	16.25
6th grade	128	12.69
7th grade	0	0.00
8th grade	0	0.00
9th grade	0	0.00
10th grade	0	0.00
11th grade	0	0.00
12th grade	0	0.00
Out of School	102	10.11

Languages	Number	%
Khmer	921	91.28
Khmer and Arabic/Cham	63	6.24
Khmer and Other	13	1.29
Arabic/Cham	7	0.69
Khmer and Vietnamese	4	0.40
Khmer, Arabic/Cham and Other	1	0.10
Missing	0	0.00
Relationship to Child	**Number**	**%**
Mother	474	46.98
Father	284	28.15
Sibling	40	3.96
Grandmother	105	10.41
Grandfather	41	4.06
Non-relative	6	0.59
Other	59	5.85
Missing	0	0.00
Caregiver Sex	**Number**	**%**
Male	361	35.78
Female	648	64.22
Missing	0	0.00

El Salvador (n=824)

Sex	Number	%
Male	419	50.85
Female	405	49.15
Missing	0	0.00
Age	**Number**	**%**
Average; SD	9.84	2.80
4	10	1.21
5	67	8.13
6	106	12.86
7	82	9.95
8	76	9.22
9	91	11.04
10	78	9.47
11	90	10.92
12	79	9.59
13	87	10.56
14	58	7.04
15+	0	0.00
Missing	0	0.00
School Year	**Number**	**%**
Kindergarten	1	0.12
1st grade	97	11.77
2nd grade	107	12.99
3rd grade	84	10.19
4th grade	57	6.92
5th grade	83	10.07
6th grade	76	9.22
7th grade	77	9.34
8th grade	50	6.07
9th grade	44	5.34
10th grade	17	2.06
11th grade	3	0.36
12th grade	0	0.00
Out of School	128	15.53

Languages	Number	%
Spanish	808	98.06
Spanish and English	12	1.46
Spanish, English and Other	2	0.24
Spanish and Nahuati-Pipil	2	0.24
Missing	0	0.00
Relationship to Child	**Number**	**%**
Mother	694	84.22
Father	71	8.62
Sibling	10	1.21
Grandmother	24	2.91
Grandfather	0	0.00
Non-relative	9	1.09
Other	16	1.94
Missing	0	0.00
Caregiver Sex	**Number**	**%**
Male	76	9.22
Female	748	90.78
Missing	0	0.00

Mozambique (n=458)

Sex	Number	%
Male	242	52.84
Female	216	47.16
Missing	0	0.00

Age	Number	%
Average; SD	13.23	2.06
4	0	0.00
5	0	0.00
6	0	0.00
7	2	0.44
8	7	1.53
9	19	4.15
10	47	10.26
11	47	10.26
12	65	14.19
13	102	22.27
14	98	21.40
15+	71	15.50
Missing	0	0.00

School Year	Number	%
Kindergarten	0	0.00
1st grade	1	0.22
2nd grade	1	0.22
3rd grade	0	0.00
4th grade	113	24.67
5th grade	17	3.71
6th grade	20	4.37
7th grade	168	36.68
8th grade	14	3.06
9th grade	4	0.87
10th grade	0	0.00
11th grade	0	0.00
12th grade	0	0.00
Out of School	120	26.20

Languages	Number	%
Emakua	249	54.37
Emakua and Português	147	32.10
Emakua and Makonde	18	3.93
Emakua and Mwani	8	1.75
Emakua and other	2	0.44
Emakua and two other languages	21	4.59
Makonde	9	1.97
Mwani	2	0.44
Português	1	0.22
Português and Makonde	1	0.22
Missing	0	0.00

Relationship to Child	Number	%
Mother	54	11.79
Father	300	65.50
Sibling	15	3.28
Grandmother	2	0.44
Grandfather	2	0.44
Non-relative	74	16.16
Other	11	2.40
Missing	0	0.00

Caregiver Sex	Number	%
Male	334	72.93
Female	124	27.07
Missing	0	0.00

Niger (n=854)

Sex	Number	%
Male	403	47.19
Female	450	52.69
Missing	1	0.12
Age	**Number**	**%**
Average; SD	10.75	2.13
4	0	0.00
5	4	0.47
6	24	2.81
7	59	6.91
8	88	10.30
9	95	11.12
10	154	18.03
11	157	18.38
12	139	16.28
13	68	7.96
14	46	5.39
15+	19	2.22
Missing	0	0.00
School Year	**Number**	**%**
Kindergarten	25	2.93
1st grade	64	7.49
2nd grade	62	7.26
3rd grade	118	13.82
4th grade	105	12.30
5th grade	102	11.94
6th grade	52	6.09
7th grade	0	0.00
8th grade	0	0.00
9th grade	0	0.00
10th grade	0	0.00
11th grade	0	0.00
12th grade	0	0.00
Out of School	326	38.17

Languages	Number	%
Hausa	176	20.61
Djarma	396	46.37
Peulh	71	8.31
Hausa and Djarma	62	7.26
Djarma and Peulh	53	6.21
Hausa and one other language	7	0.82
Hausa and two other languages	21	2.46
Djarma and one other language	28	3.28
Djarma and two other languages	2	0.23
French	3	0.35
French and one other language	2	0.23
Other	20	2.34
Relationship to Child	**Number**	**%**
Mother	385	45.08
Father	221	25.88
Sibling	55	6.44
Grandmother	50	5.85
Grandfather	107	12.53
Non-relative	16	1.87
Other	19	2.22
Missing	0	0.00
Caregiver Sex	**Number**	**%**
Male	319	37.35
Female	534	62.53
Missing	0	0.00

Occupied Palestinian Territories (n=1142)

Sex	Number	%
Male	565	49.47
Female	577	50.53
Missing	0	0.00
Age	**Number**	**%**
Average; SD	10.04	2.38
4	0	0.00
5	44	3.85
6	95	8.32
7	116	10.16
8	142	12.43
9	159	13.92
10	152	13.31
11	164	14.36
12	124	10.86
13	98	8.58
14	49	4.29
15+	0	0.00
Missing	0	0.00
School Year	**Number**	**%**
Kindergarten	0	0.00
1st grade	52	4.55
2nd grade	96	8.41
3rd grade	123	10.77
4th grade	155	13.57
5th grade	147	12.87
6th grade	156	13.66
7th grade	162	14.19
8th grade	116	10.16
9th grade	97	8.49
10th grade	25	2.19
11th grade	0	0.00
12th grade	0	0.00
Out of School	23	2.01

Languages	Number	%
Only Arabic	1020	89.32
Arabic and English	84	7.36
Arabic and one other language	26	2.28
Arabic and two other languages	8	0.70
Arabic and three other languages	2	0.18
Arabic and four other languages	1	0.09
English	1	0.09
Missing	0	0.00
Relationship to Child	**Number**	**%**
Mother	968	84.76
Father	123	10.77
Sibling	31	2.71
Grandmother	3	0.26
Grandfather	3	0.26
Non-relative	1	0.09
Other	13	1.14
Missing	0	0.00
Caregiver Sex	**Number**	**%**
Male	193	16.90
Female	949	83.10
Missing	0	0.00

Philippines (n=798)

Sex	Number	%
Male	367	45.99
Female	431	54.01
Missing	0	0.00

Age	Number	%
Average; SD	12.68	1.80
4	0	0.00
5	0	0.00
6	0	0.00
7	0	0.00
8	4	0.50
9	25	3.13
10	150	18.80
11	143	17.92
12	120	15.04
13	135	16.92
14	129	16.17
15+	92	11.53
Missing	0	0.00

School Year	Number	%
Kindergarten	0	0.00
1st grade	0	0.00
2nd grade	1	0.13
3rd grade	0	0.00
4th grade	11	1.38
5th grade	51	6.39
6th grade	112	14.04
7th grade	148	18.55
8th grade	136	17.04
9th grade	101	12.66
10th grade	99	12.41
11th grade	77	9.65
12th–16th grade	17	2.13
Out of School	45	5.64

Relationship to Child	Number	%
Mother	561	70.30
Father	112	14.04
Sibling	28	3.51
Grandmother	43	5.39
Grandfather	1	0.13
Non-relative	35	4.39
Other	18	2.26
Missing	0	0.00

Caregiver Sex	Number	%
Male	135	16.92
Female	663	83.08
Missing	0	0.00

Sudan (n=559)

Sex	Number	%
Male	194	34.70
Female	365	65.30
Missing	1	0.18
Age	**Number**	**%**
Average; SD	11.44	2.62
4	0	0.00
5	4	0.64
6	23	3.68
7	39	6.24
8	48	7.68
9	65	10.40
10	82	13.12
11	97	15.52
12	72	11.52
13	70	11.20
14	59	9.44
15+	65	10.40
Missing	1	0.16
School Year	**Number**	**%**
Kindergarten	0	0.00
1st grade	14	2.50
2nd grade	17	3.04
3rd grade	6	1.07
4th grade	2	0.36
5th grade	9	1.61
6th grade	18	3.22
7th grade	1	0.18
8th grade	0	0.00
9th grade	0	0.00
10th grade	0	0.00
11th grade	0	0.00
12th grade	0	0.00
Out of School	492	88.01

Languages	Number	%
Arabic Rashad	345	61.72
Arabic and Tagli	77	13.77
Arabic and other	58	10.38
Arabic and Bargo	32	5.72
Arabic and Kawalib	12	2.15
English	12	2.15
Arabic and one other language	12	2.15
Arabic and two other languages	6	1.07
English and one other language	2	0.36
Tagli	2	0.36
Bargo	1	0.18
Caregiver Sex	**Number**	**%**
Male	277	49.55
Female	282	50.45
Missing	0	0.00

APPENDIX B: IRR BY SUB-DOMAIN

El Salvador (n=824)

ReAL Sub-Domain	Percent Agreement (%)						
	oPt	Mozambique	Philippines	Cambodia	Sudan	Niger	El Salvador**
	High Access	High Access	High Access	High Access	High Access	High Access	Caregiver Report
	(n=185)	(n=79)	(n=92)	(n=208)	(n=355)	(n=219)	(n=208)
Literacy	**98.92**	**93.07**	**97.04**	**98.34**	**93.15**	**97.62**	**97.47**
Expressive Vocabulary	98.37	83.75	95.11	96.64	87.76	97.49	99.75
Retelling Story	97.85	94.87	95.66	98.56	85.72	99.32	97.00
Listening Comprehension	98.17	92.31	97.83	99.23	94.70	98.72	98.75
Letter Identification	99.14	95.32	98.78	97.77	95.41	96.90	97.00
Common Word Identification	99.25	90.90	100.00	98.42	95.51	97.53	96.67
Sentence Comprehension	98.38	94.61	95.29	99.14	88.78	97.17	98.00
Reading Comprehension	99.00	91.19	95.65	98.90	90.24	98.63	94.00
Numeracy	**99.06**	**94.03**	**96.42**	**97.40**	**88.40**	**94.46**	**98.35**
One to One Correspondence	99.10	99.57	N/A*	99.36	93.88	95.89	98.50
Number Identification	99.46	96.80	96.54	98.44	93.71	94.14	97.63
Addition	98.49	91.41	97.50	94.67	91.13	93.38	99.50
Subtraction	99.19	92.69	95.44	97.84	76.43	94.93	99.00
Word Problems	98.92	90.60	95.65	98.88	92.52	96.35	97.75
Social Emotional Learning	**97.27**	**90.62**	**92.02**	**93.17**	**86.41**	**94.17**	**94.17**
Relationships	98.06	90.68	92.03	97.98	87.48	93.94	94.57
Stress Management	96.77	93.16	90.95	96.31	88.09	95.28	96.00
Empathy	97.10	91.16	93.70	96.88	86.94	93.47	96.80
Conflict Resolution	96.10	92.63	89.13	97.72	85.71	96.12	98.13
Self-Concept	96.60	86.97	91.67	98.00	82.48	94.07	98.08

*No variation in responses, every child answered this question correctly in the Philippines so IRR could not be calculated

**El Salvador used the caregiver reported modality, as such there are fewer questions within each sub-domain and overall domain for literacy and numeracy, SEL questions are the same

APPENDIX C: CFA FIT STATISTICS BY COUNTRY

Cambodia

Subdomain	$\chi 2$(df)	*p*	RMSEA [90%CI]	CFI	SRMR
Literacy					
Listening comprehension	4.481 (5)	.482	.024 [.000, .202]	.999	.028
Letter/letter sound identification	259.493 (170)	< .001	.023 [.017, .028]	.999	.046
Common word identification	37.159 (35)	.370	.008 [.000, .026]	1.000	.017
Sentence–level comprehension			Heywood		
Oral passage reading	12.868 (14)	0.537	.000 [.000, .154]	1.000	.063
Numeracy					
Number identification	112.531 (54)	< .001	.033 [.025, .042]	.999	.040
Addition	103.278 (35)	< .001	.047 [.037, .058]	.999	.040
Subtraction	62.130 (35)	.003	.030 [.017, .041]	1.000	.038
Social Emotional Learning					
Self-concept[1]			Heywood		
Use of social supports[2]	0.423 (2)	.810	.000 [.000, .000]	1.000	.138
Help seeking behavior[3]	0.752 (2)	.687	.000 [.000, .048]	1.000	.134
Stress management[4]			covariance matrix of latent variables is not positive definite		
Empathy[5]			Heywood		
Conflict resolution[6]			Heywood		

Note
Robust RMSEA and robust CFI were reported for listening comprehension, oral passage reading, and use of social support based on Savalei (2021).

1
Six items (sel1–sel6) loaded on one latent factor. The tetrachoric correlations between items ranged from .97 to 1.00.

2
Four items (rel3, rel4, rel9, and rel13) loaded on one latent factor. The tetrachoric correlation coefficient between rel3 and rel4 was .99.

3
Four items (rel5, rel6, rel10, and rel14) loaded on the latent factor. The tetrachoric correlation coefficient between rel5 and rel6 was .99.

4
Four items (rel1, rel2, rel8, and rel12) loaded on social support, and three items loaded on behavioral regulation (st1–st3). The tetrachoric correlation coefficient between rel1 and rel2 was .98. The tetrachoric correlation coefficients between st1, st2, and st3 ranged from .96 to .98.

5
Five items (rel7, rel11, rel15, e1, and e6) loaded on identifying feelings of others, and six items (e2, e3, e5, e7, e8, and e10) loaded on empathy. The tetrachoric correlation coefficients between e1, e2, and e3 were .98. The tetrachoric correlation coefficient between e6 and e7 was .97. The tetrachoric correlation coefficient between e6 and e8 was .95.

6
Four items (con1–con4) loaded on social problem solving, and two items (e4 and e9) loaded on interpreting hostility. The tetrachoric correlation coefficient between con1 and other social problem solving items were greater than .96.

Mozambique

Subdomain	$\chi 2$(df)	*p*	RMSEA [90%CI]	CFI	SRMR
Literacy					
Listening comprehension	7.752 (5)	.170	.186 [.000, .444]	.985	.009
Letter/letter sound identification	191.790 (170)	.121	.017 [.000, .028]	1.000	.037
Common word identification	51.842 (35)	.033	.0037 [.011, .057]	1.000	.007
Sentence–level comprehension			Heywood		
Oral passage reading			Failed to converge		
Numeracy					
Number identification	57.257 (54)	.355	.012 [.000, .032]	1.000	.045
Addition	70.849 (35)	< .001	.049 [.033, .066]	.998	.032
Subtraction	150.316 (35)	< .001	.090 [.075, .105]	.995	.061
Social Emotional Learning					
Self-concept[1]	4.598 (9)	.868	.000 [.000, .028]	1.000	.007
Use of social supports[2]	0.639 (2)	.726	.000 [.000, .066]	1.000	.131
Help seeking behavior[3]	0.482 (2)	.786	.000 [.000, .060]	1.000	.129
Stress management[4]			Heywood		
Empathy[5]			Heywood		
Conflict resolution[6]			Heywood		

Note
Robust RMSEA and robust CFI were reported for listening comprehension based on Savalei (2021).

1
Six items (sel1–sel6) loaded on one latent factor. The tetrachoric correlations between items ranged from .97 to 1.00.

2
Four items (rel3, rel4, rel9, and rel13) loaded on one latent factor. The tetrachoric correlation coefficient between rel3 and rel4 was .99.

3
Four items (rel5, rel6, rel10, and rel14) loaded on the latent factor. The tetrachoric correlation coefficient between rel5 and rel6 was .99.

4
Four items (rel1, rel2, rel8, and rel12) loaded on social support, and three items loaded on behavioral regulation (st1–st3). The tetrachoric correlation coefficient between rel1 and rel2 was .98. The tetrachoric correlation coefficients between st1, st2, and st3 ranged from .96 to .98.

5
Five items (rel7, rel11, rel15, e1, and e6) loaded on identifying feelings of others, and six items (e2, e3, e5, e7, e8, and e10) loaded on empathy. The tetrachoric correlation coefficients between e1, e2, and e3 were .98. The tetrachoric correlation coefficient between e6 and e7 was .97. The tetrachoric correlation coefficient between e6 and e8 was .95.

6
Four items (con1–con4) loaded on social problem solving, and two items (e4 and e9) loaded on interpreting hostility. The tetrachoric correlation coefficient between con1 and other social problem solving items were greater than .96.

Niger

Subdomain	$\chi 2$(df)	*p*	RMSEA [90%CI]	CFI	SRMR
Literacy					
Listening comprehension	3.432 (5)	.634	.000 [.000, .159]	1.00	.022
Letter/letter sound identification	374.397 (170)	< .001	.038 [.033, .043]	.981	.067
Common word identification	110.885 (35)	< .001	.181 [.144, .219]	.901	.045
Sentence–level comprehension	114.610 (5)	< .001	.363 [.291, .439]	.735	.119
Oral passage reading	33.447 (14)	.002	.097 [.000, .177]	.975	.051
Numeracy					
Number identification	163.420 (54)	< .001	.049 [.041, .058]	.994	.058
Addition	86.226 (35)	< .001	.123 [.089, .158]	.958	.028
Subtraction	80.846 (35)	< .001	.142 [.101, .183]	.942	.033
Social Emotional Learning					
Self-concept[1]			Heywood		
Use of social supports[2]	0.827 (2)	.661	.000 [.000, .053]	1.000	.190
Help seeking behavior[3]	0.326 (2)	.850	.000 [.000, .000]	1.000	.151
Stress management[4]			covariance matrix of latent variables is not positive definite		
Empathy[5]			Heywood		
Conflict resolution[6]			Heywood		

Note
Robust RMSEA and robust CFI were reported for listening comprehension, common word identification, sentence-level comprehension, oral passage reading, addition, subtraction, and help seeking behavior based on Savalei (2021).

1
Six items (sel1–sel6) loaded on one latent factor. The tetrachoric correlations between items ranged from .94 to .99.

2
Four items (rel3, rel4, rel9, and rel13) loaded on one latent factor. The tetrachoric correlation coefficient between rel3 and rel4 was .99.

3
Four items (rel5, rel6, rel10, and rel14) loaded on the latent factor. The tetrachoric correlation coefficient between rel5 and rel6 was .99.

4
Four items (rel1, rel2, rel8, and rel12) loaded on social support, and three items loaded on behavioral regulation (st1–st3). The tetrachoric correlation coefficient between rel1 and rel2 was .97. The tetrachoric correlation coefficients between st1, st2, and st3 ranged from .95 to .98.

5
Five items (rel7, rel11, rel15, e1, and e6) loaded on identifying feelings of others, and six items (e2, e3, e5, e7, e8, and e10) loaded on empathy. The tetrachoric correlation coefficients between e1, e2, and e3 range from .92–.98. The tetrachoric correlation coefficient between e6 and e7 was .95. The tetrachoric correlation coefficient between e6 and e8 was .95.

6
Four items (con1–con4) loaded on social problem solving, and two items (e4 and e9) loaded on interpreting hostility. The tetrachoric correlation coefficient between con1 and other social problem solving items were greater than .95.

oPt

Subdomain	$\chi 2$(df)	*p*	RMSEA [90%CI]	CFI	SRMR
Literacy					
Listening comprehension	0.589 (5)	.988	.000 [.000, .000]	1.000	.015
Letter/letter sound identification	163.488 (170)	.626	.000 [.000, .012]	1.000	.043
Common word identification	46.454 (35)	.093	.160 [.101, .217]	.903	.053
Sentence–level comprehension	60.410 (5)	< .001	.698 [.467, .952]	.483	.178
Oral passage reading	11.067 (14)	0.681	.000 [.000, .108]	1.000	.047
Numeracy					
Number identification	87.181 (54)	.003	.024 [.014, .032]	.991	.097
Addition	48.225 (35)	.068	.101 [.033, .156]	.950	.058
Subtraction	70.745 (35)	< .001	.133 [.078, .184]	.902	.077
Social Emotional Learning					
Self-concept[1]	68.174 (9)	< .001	.078 [.061, .096]	.975	.051
Use of social supports[2]	0.242 (2)	.886	.000 [.000, .028]	1.000	.173
Help seeking behavior[3]	0.442 (2)	.802	.000 [.000, .037]	1.000	.227
Stress management[4]			Heywood		
Empathy[5]			Heywood		
Conflict resolution[6]			Heywood		

Note

Robust RMSEA and robust CFI were reported for listening comprehension, common word identification, sentence-level comprehension, oral passage reading, addition, and subtraction based on Savalei (2021).

1 Six items (sel1–sel6) loaded on one latent factor. The tetrachoric correlations between items ranged from .48 to .94.

2 Four items (rel3, rel4, rel9, and rel13) loaded on one latent factor. The tetrachoric correlation coefficient between rel3 and rel4 was .97.

3 Four items (rel5, rel6, rel10, and rel14) loaded on the latent factor. The tetrachoric correlation coefficient between rel5 and rel6 was .97.

4 Four items (rel1, rel2, rel8, and rel12) loaded on social support, and three items loaded on behavioral regulation (st1–st3). The tetrachoric correlation coefficient between rel1 and rel2 was .97. The tetrachoric correlation coefficients between st1, st2, and st3 ranged from .97 to .98.

5 Five items (rel7, rel11, rel15, e1, and e6) loaded on identifying feelings of others, and six items (e2, e3, e5, e7, e8, and e10) loaded on empathy.

6 Four items (con1–con4) loaded on social problem solving, and two items (e4 and e9) loaded on interpreting hostility. The tetrachoric correlation coefficient between con1 and other social problem solving items ranged from .93 to .97.

Philippines

Subdomain	$\chi 2$(df)	*p*	RMSEA [90%CI]	CFI	SRMR
Literacy					
Listening comprehension	7.668 (5)	.175	.097 [.000, .318]	.943	.089
Letter/letter sound identification			Heywood		
Common word identification	85.521 (35)	< .001	.043 [.031, .054]	0.963	.130
Sentence–level comprehension			Heywood		
Oral passage reading	53.216 (14)	< .001	.188 [.130, .250]	.704	.107
Numeracy					
Number identification	100.462 (54)	< .001	.033 [.023, .043]	.994	.080
Addition	47.874 (35)	.072	.250 [.155, .341]	.688	.122
Subtraction	46.200 (35)	.098	.254 [.184, .325]	.694	.115
Social Emotional Learning					
Self-concept[1]			Heywood		
Use of social supports[2]	0.432 (2)	.806	.000 [.000, .000]	1.000	.169
Help seeking behavior[3]	18.472 (2)	.000	.317 [.128, .540]	.931	.080
Stress management[4]			Heywood		
Empathy[5]			Sample covariance matrix is not positive–definite		
Conflict resolution[6]			Heywood		

Note
Robust RMSEA and robust CFI were reported for listening comprehension, oral passage reading, addition, subtraction, use of social supports, and help seeking behavior based on Savalei (2021).

1
Six items (sel1–sel6) loaded on one latent factor. The tetrachoric correlations between items ranged from .94 to .99.

2
Four items (rel3, rel4, rel9, and rel13) loaded on one latent factor. The tetrachoric correlation coefficient between rel3 and rel4 was .97.

3
Four items (rel5, rel6, rel10, and rel14) loaded on the latent factor. The tetrachoric correlation coefficient between rel5 and rel6 was .95.

4
Four items (rel1, rel2, rel8, and rel12) loaded on social support, and three items loaded on behavioral regulation (st1–st3). The tetrachoric correlation coefficient between rel1 and rel2 was .95. The tetrachoric correlation coefficients between st1, st2, and st3 ranged from .95 to .97.

5
Five items (rel7, rel11, rel15, e1, and e6) loaded on identifying feelings of others, and six items (e2, e3, e5, e7, e8, and e10) loaded on empathy. The tetrachoric correlation coefficients among e1, e2, and e3 ranged from .95 to .99. The tetrachoric correlation coefficient between e6 and e7 was .92. The tetrachoric correlation coefficient between e6 and e8 was .91.

6
Four items (con1–con4) loaded on social problem solving, and two items (e4 and e9) loaded on interpreting hostility. The tetrachoric correlation coefficient between con1 and other social problem solving items ranged from .89 to .95.

Sudan

Subdomain	$\chi 2$(df)	*p*	RMSEA [90%CI]	CFI	SRMR
Literacy					
Listening comprehension	3.815 (5)	.576	.000 [.000, .225]	1.000	.034
Letter/letter sound identification	246.896 (170)	< .001	.028 [.020, .035]	.991	.069
Common word identification	68.620 (35)	.001	.041 [.026, .055]	.993	.017
Sentence–level comprehension	99.936 (5)	< .001	.629 [.483, .787]	.602	.143
Oral passage reading	18.461 (14)	0.187	.043 [.000, .091]	.997	.058
Numeracy					
Number identification	114.343 (54)	< .001	.044 [.033, .055]	.989	.074
Addition	52.387 (35)	.030	.076 [.000, .153]	.988	.026
Subtraction	88.217 (35)	< .001	.233 [.165, .301]	.878	.051
Social Emotional Learning					
Self-concept[1]			Heywood		
Use of social supports[2]			Heywood		
Help seeking behavior[3]	0.444 (2)	.801	.000 [.000, .053]	1.000	.124
Stress management[4]			Heywood		
Empathy[5]			Heywood		
Conflict resolution[6]			Heywood		

Note
Robust RMSEA and robust CFI were reported for listening comprehension, sentence-level comprehension, oral passage reading, addition, and subtraction based on Savalei (2021).

1
Six items (sel1–sel6) loaded on one latent factor. The tetrachoric correlations between items ranged from .67 to .99.

2
Four items (rel3, rel4, rel9, and rel13) loaded on one latent factor. The tetrachoric correlation coefficient between rel3 and rel4 was .97.

3
Four items (rel5, rel6, rel10, and rel14) loaded on the latent factor. The tetrachoric correlation coefficient between rel5 and rel6 was .99.

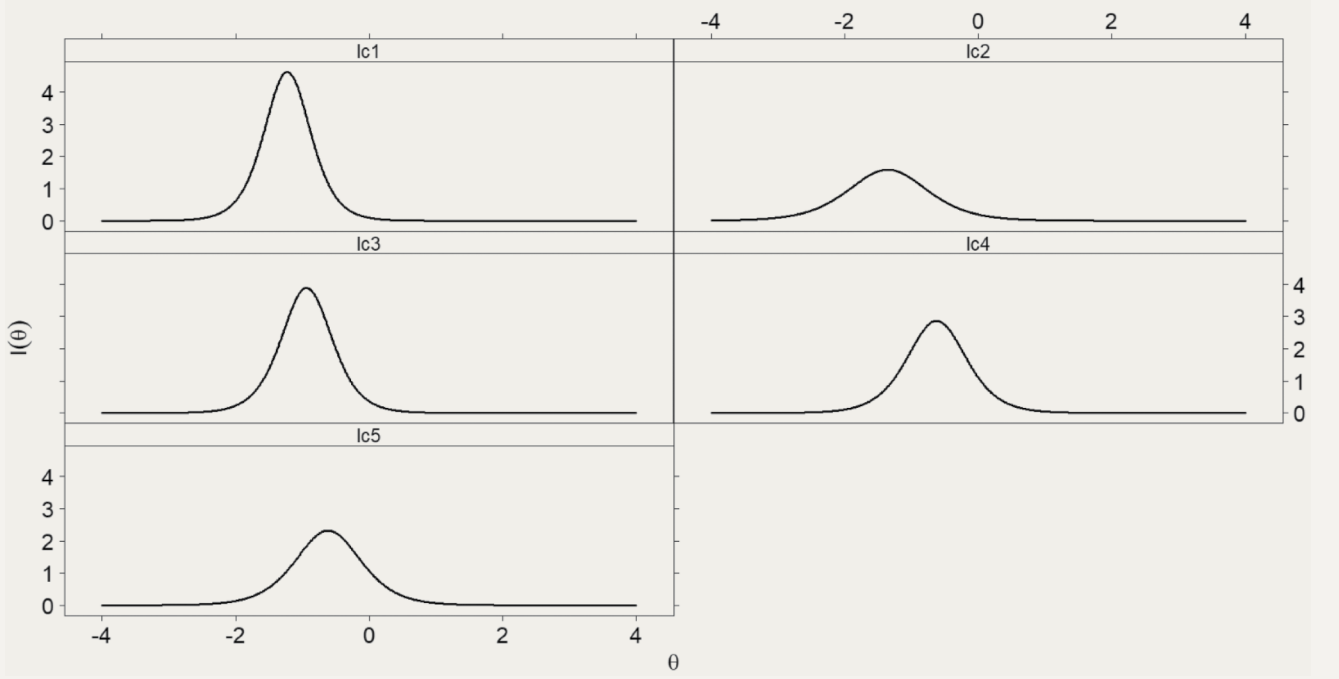
4
Four items (rel1, rel2, rel8, and rel12) loaded on social support, and three items loaded on behavioral regulation (st1–st3). The tetrachoric correlation coefficient between rel1 and rel2 was .97. The tetrachoric correlation coefficients between st1, st2, and st3 ranged from .95 to .97.

5
Five items (rel7, rel11, rel15, e1, and e6) loaded on identifying feelings of others, and six items (e2, e3, e5, e7, e8, and e10) loaded on empathy. The tetrachoric correlation coefficients among e1, e2, and e3 ranged from .96 to .98. The tetrachoric correlation coefficient between e6 and e7 was .98. The tetrachoric correlation coefficient between e6 and e8 was .98.

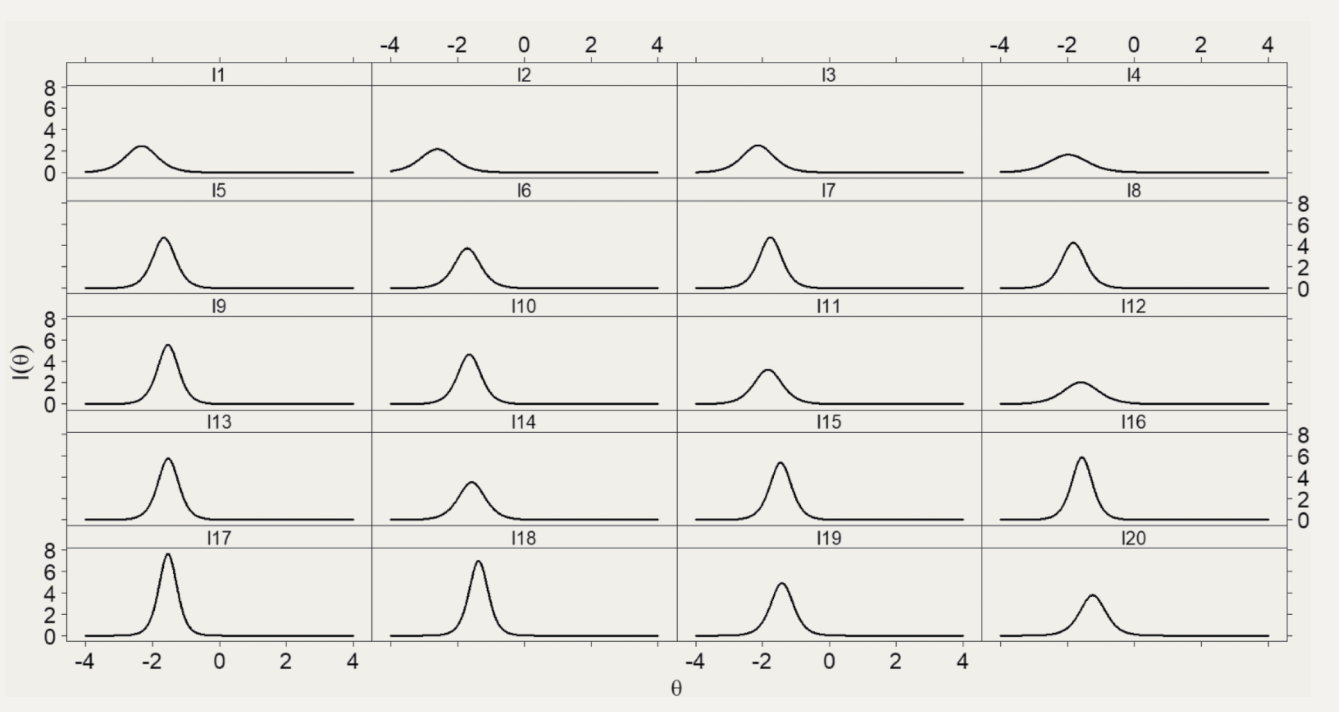
6
Four items (con1–con4) loaded on social problem solving, and two items (e4 and e9) loaded on interpreting hostility. The tetrachoric correlation coefficient between con1 and other social problem solving items ranged from .87 to .97.

APPENDIX D: ITEM INFORMATION FUNCTIONS

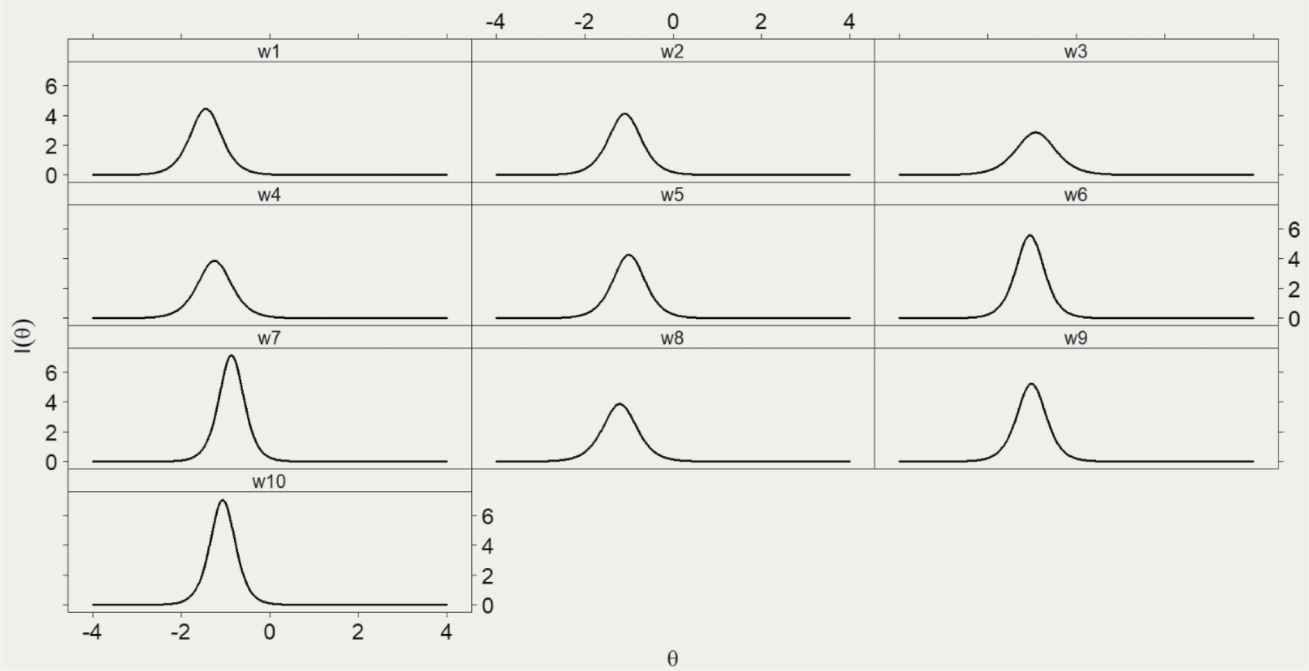
Item Information Function for Listening Comprehension



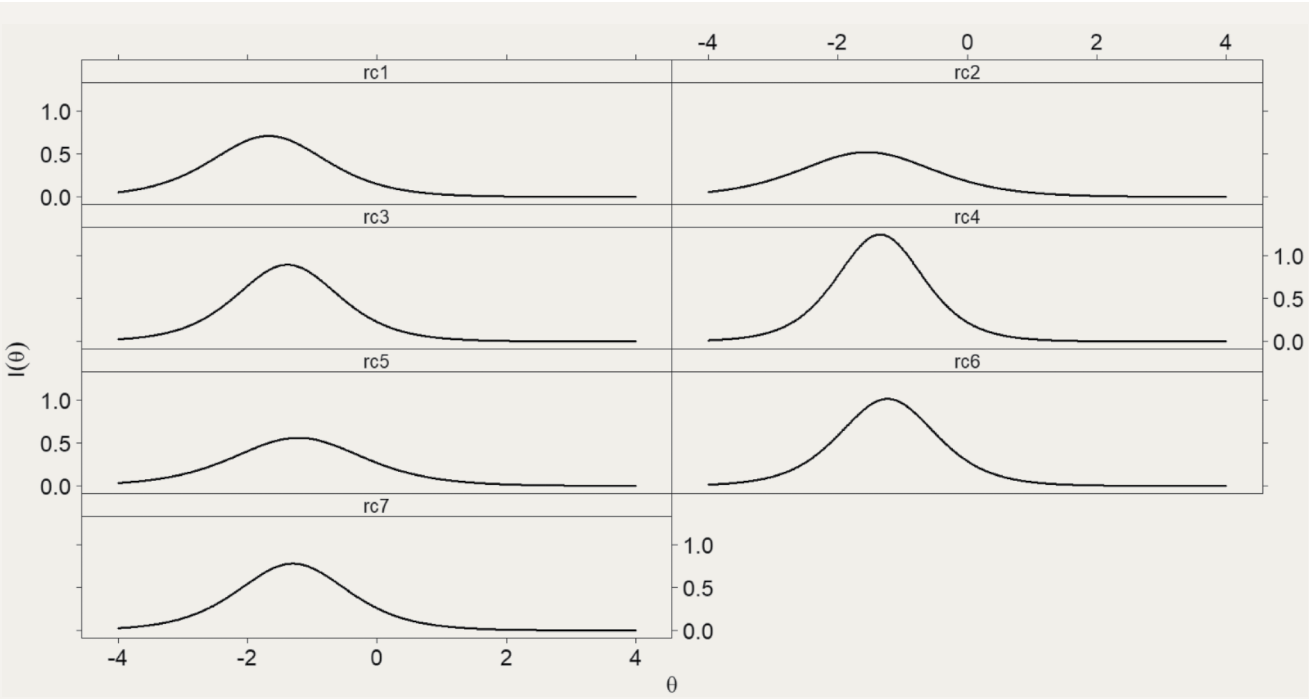
Item Information Function for Letter/Letter Sound Identification



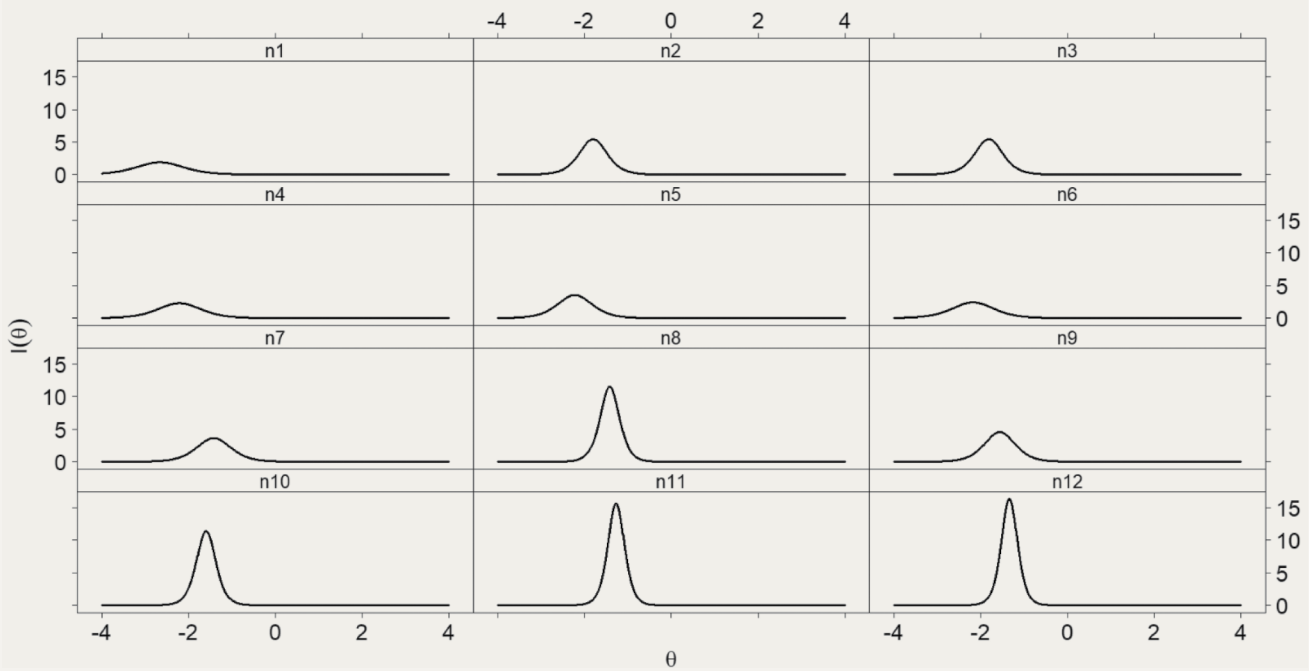
Item Information Function for Common Word Identification



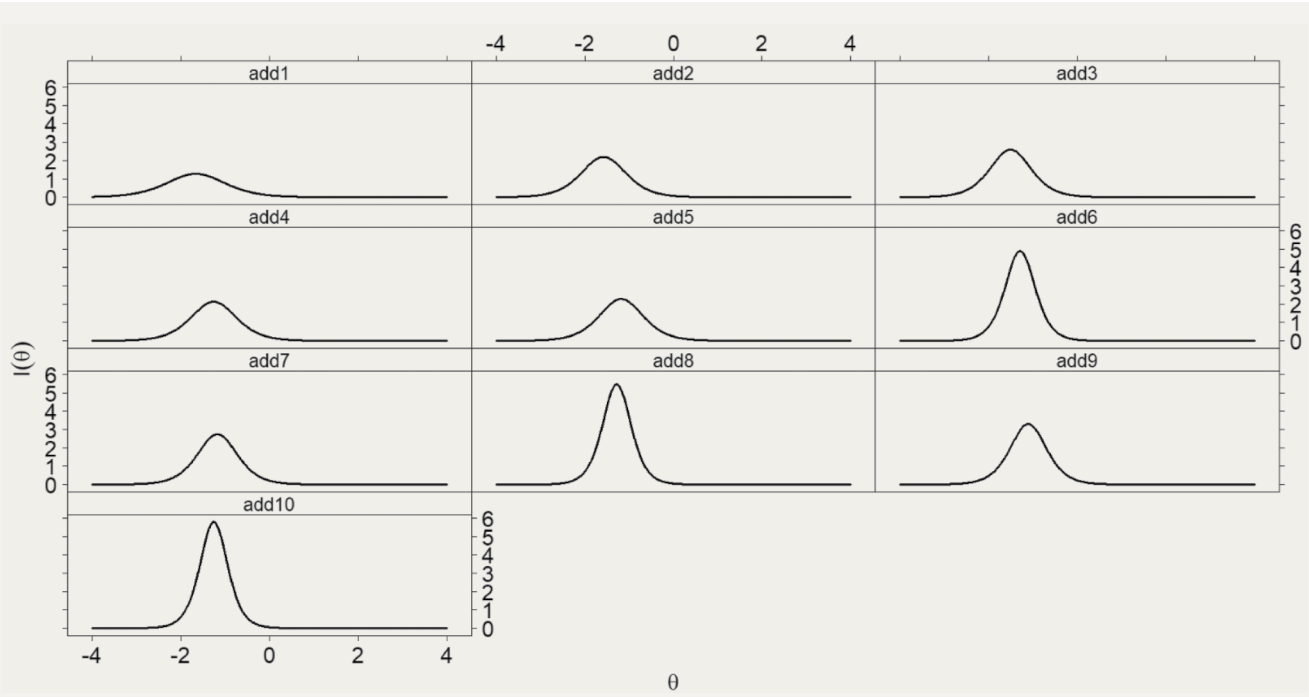
Item Information Function for Oral Passage Reading Comprehension



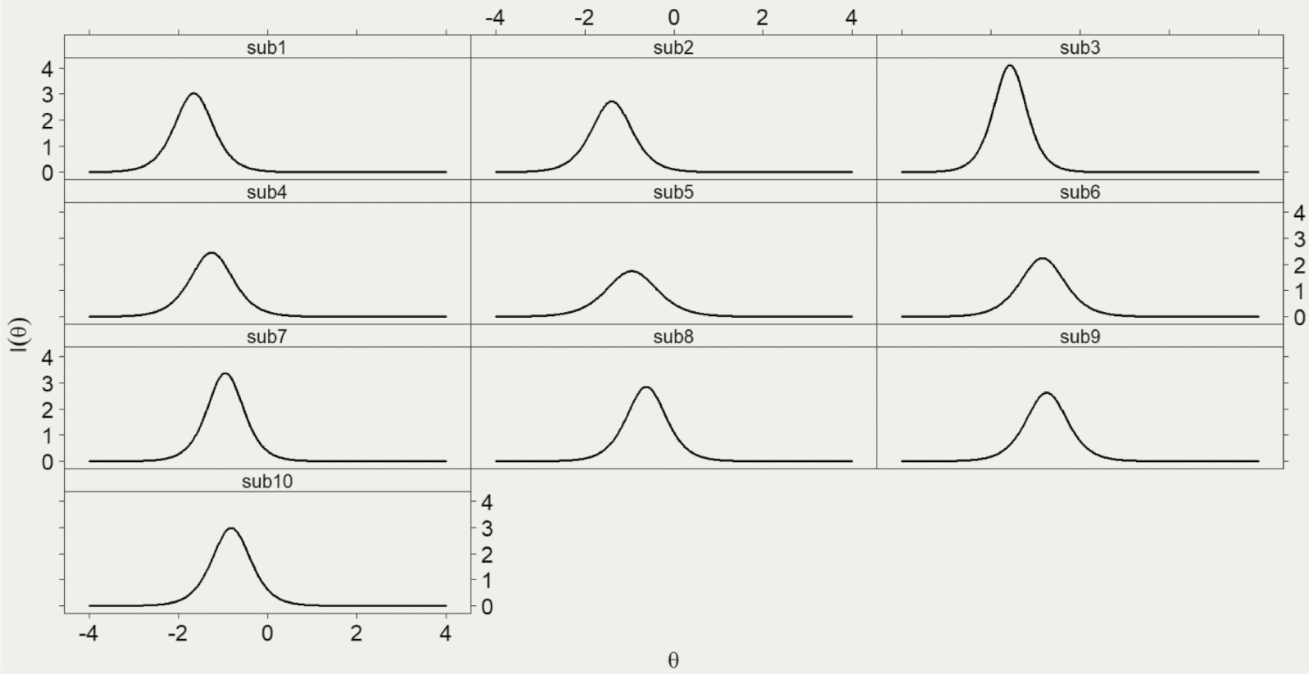
Item Information Function for Number Identification



Item Information Function for Addition

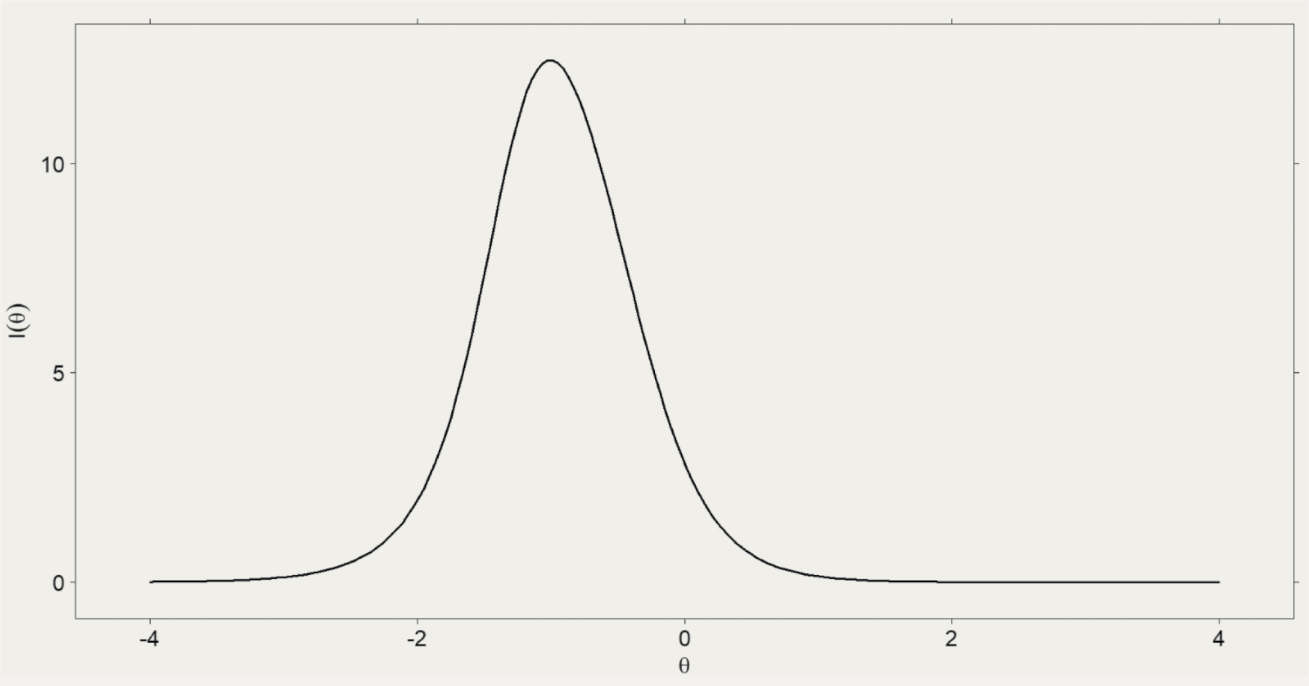


Item Information Function for Subtraction

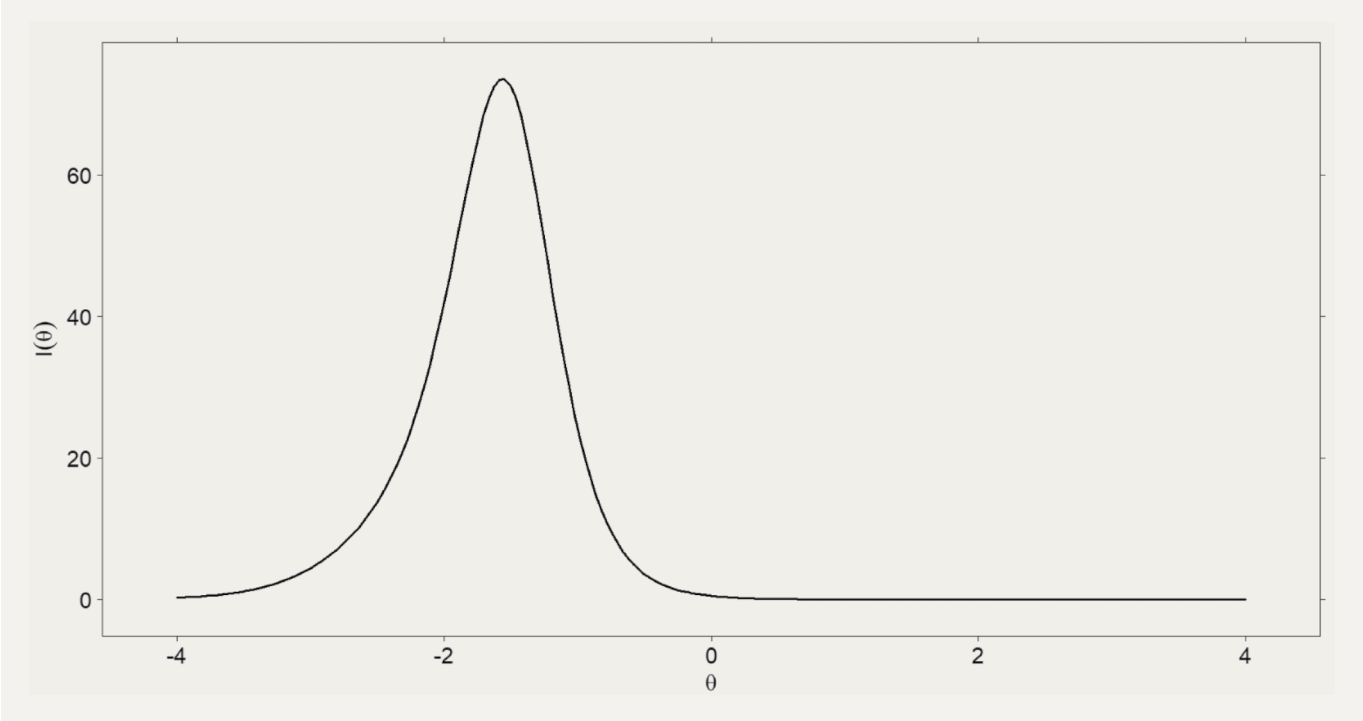


APPENDIX E: TEST INFORMATION FUNCTIONS

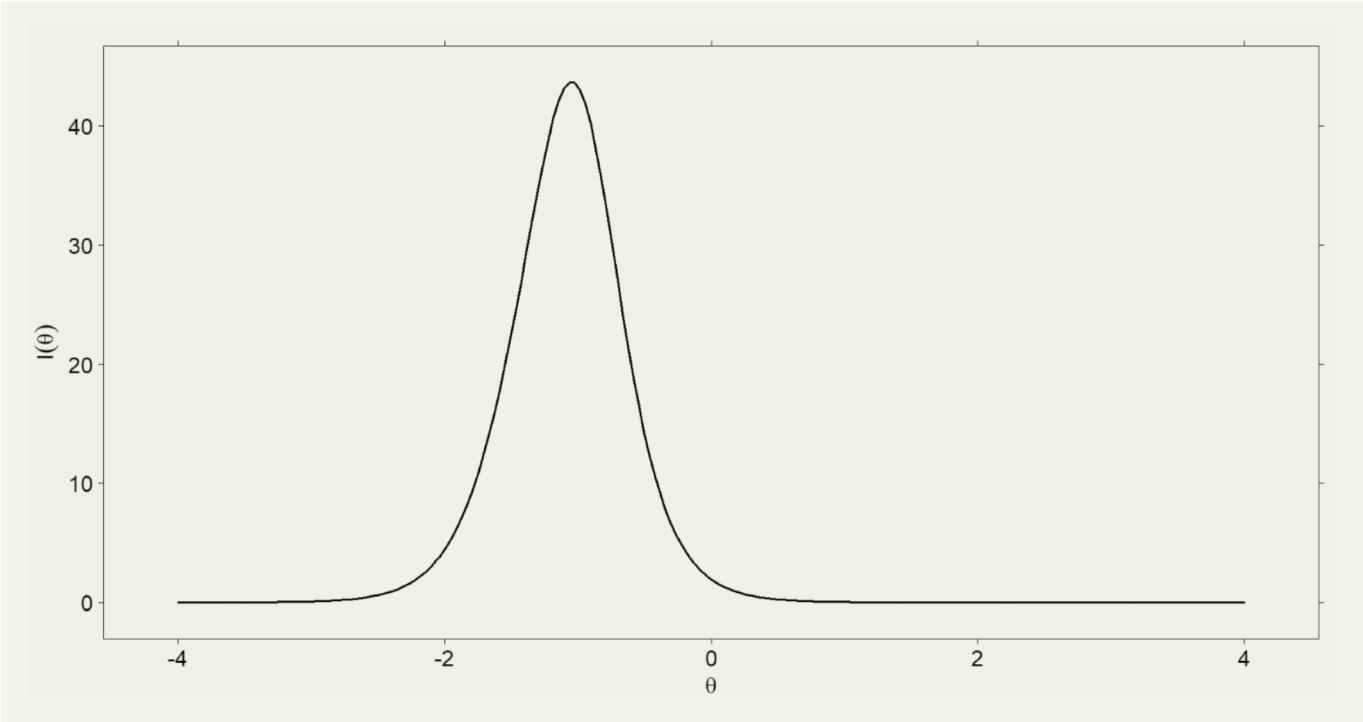
Test Information Functioning for Listening Comprehension



Test Information Functioning for Letter/Letter Sound Identification

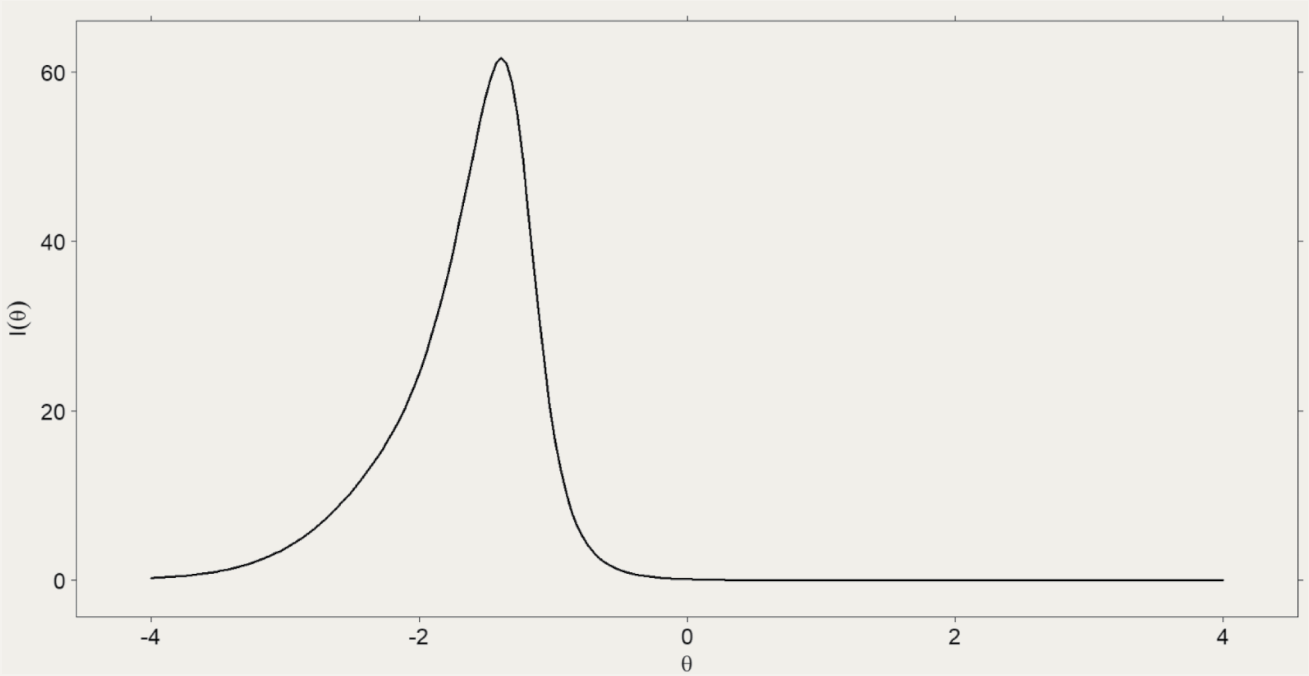


Test Information Functioning for Common Word Identification

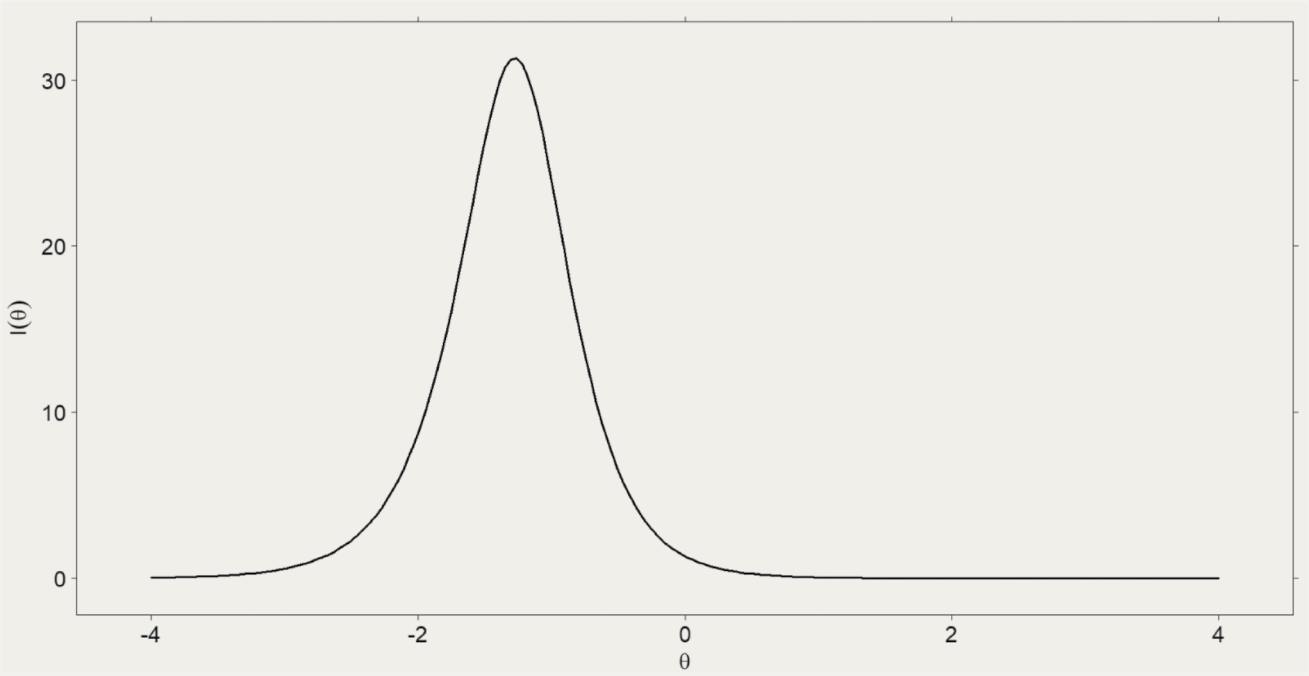


Test Information Function for Oral Passage Reading

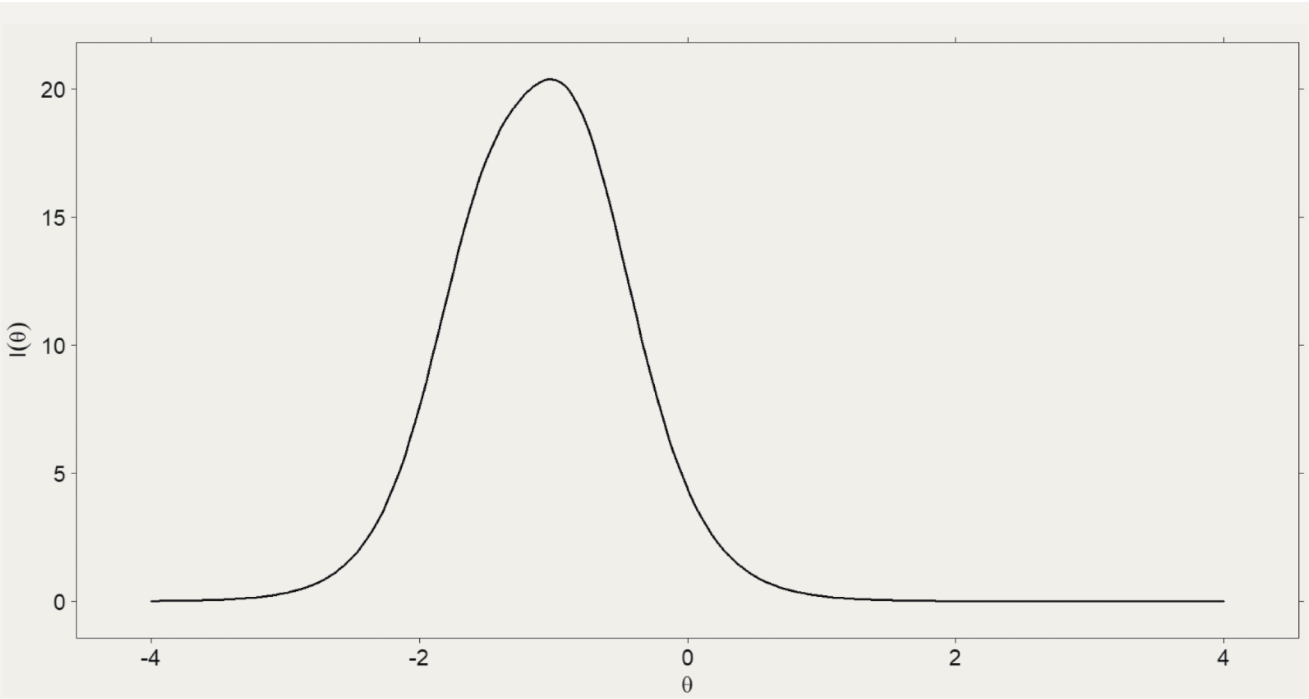
Test Information Functioning for Number Identification



Test Information Functioning for Addition



Test Information Functioning for Subtraction



APPENDIX F: REAL HIGH ACCESS ADMINISTRATION GUIDANCE

See link: https://resourcecentre.savethechildren.net/pdf/ReAL-High-English.pdf/

APPENDIX G: REAL CAREGIVER REPORT ADMINISTRATION GUIDANCE

See link: https://resourcecentre.savethechildren.net/pdf/ReAL-Caregiver-English.pdf

APPENDIX H: QUALITATIVE DATA COLLECTION PROTOCOLS

Please describe your role in the ReAL beta phase.

Section 1: Perceptions Around Feasibility

Technical Infrastructure:

1a Did the team and participants have the necessary technical infrastructure (e.g., stable internet and devices) to implement the remote learning assessment tool effectively?

1b What challenges, if any, did the data collection team encounter with the technology?

Logistical Challenges:

2a What logistical challenges, if any, (e.g., coordinating with participants, data collection, accessing participant contact information) impacted the feasibility of using the tool?

2b Did the team take any steps to address these challenges? If so, how were they resolved?

Assessment Environment:

3 How did the assessment environment (example. location, presence of a caregiver, communication modality) differ from similar types of in-person assessments in your context, and how could this have affected the responses of children?

Staff Training and Implementation:

4a What enumerator skill-related challenges, if any, did the team encounter when implementing the tool?

4b Did the team take any steps to address these challenges? If so, how were they resolved?

Scalability:

5 Based on your experience, do you think this tool could be scaled up for use in your country? What target populations and factors would need to be considered for scaling?

Section II: Perceptions Around Appropriateness

Contextualization and Engagement:

6 After tailoring the test materials and stimuli to align with the local context—by developing context-specific stories and stimuli, and reviewing the translated versions—how effectively did these materials resonate with and engage the different target groups (e.g., in-school vs. out-of-school, rural vs. urban, language variations)

7a What specific challenges or gaps in children's and caregivers' understanding of the questions, if any, did the data collection team observe among participants?

7b Were there particular groups or contexts where these issues were more pronounced?

Content Relevance:

8a To what extent does the team believe this tool aligns with and reflects the educational goals and curriculum of your country context?

8b Are there specific aspects where the tool is particularly effective or areas where it falls short?

Design and Suitability:

9 Did the team receive feedback from communities before and/or after data collection? If yes, what feedback was received?

Section III: Additional Questions

10 How satisfied is the team with the remote learning assessment tool in terms of its feasibility and appropriateness?

11 What would the team change, if anything, to improve the tool's fit for your context?

12 What advice would the team give to other COs in similar contexts who are considering using this tool?

13 Is there anything else you would like to add (examples, lessons learned, etc.)?

Save the Children