Scaling Education Innovations in Complex Emergencies

Evidence From the Humanitarian Education Accelerator

OCTOBER 2019

Thomas de Hoop | Hannah Ring | Andrea Coombes | Victoria Rothbard
Chinmaya Holla | Kelsey Hunt | David Seidenfeld | Helen Connolly

Produced by AIR with support from UNICEF and UNHCR and funding from DFID

AIR®
AMERICAN INSTITUTES FOR RESEARCH®

unicef

UNHCR
The UN Refugee Agency

UKaid
from the British people

Scaling Education Innovations in Complex Emergencies

Evidence From the Humanitarian Education Accelerator

OCTOBER 2019

Thomas de Hoop | Hannah Ring | Andrea Coombes | Victoria Rothbard
Chinmaya Holla | Kelsey Hunt | David Seidenfeld | Helen Connolly

Produced by AIR with support from UNICEF and UNHCR and funding from DFID

9619_08/19

Acknowledgements

We recognise the contributions of many organisations, without which it would not have been possible to complete this report on the meta-evaluation for the Humanitarian Education Accelerator (HEA). Our thanks go to the United Kingdom's Department for International Development, the United Nations Children's Fund, and the United Nations High Commissioner for Refugees for the opportunity to carry out this study, and for their financial and technical support. Kepler, War Child Holland, the World University Service of Canada, Libraries Without Borders, and Caritas Switzerland also made important technical contributions to this report. Finally, we acknowledge the members of the Reference Group for the HEA, whose inputs guided the direction of this report.

Many individuals also made important contributions and are listed below in alphabetical order:

> Felipe Barrera-Osorio, Chris Berry, Matt Brossard, Felicity Brown, Serafina Buzby, Mohammed El Desouky, Thomas Dreesen, Ian Gray, Ashley Haywood, Henri van den Idsert, Darius Isaboke, Charlotte Jenner, Andrea Jetten, Mark Jordans, Kathleen Letshabo, Timothy Kinoti, Naomi Maina, Laura Miller, Alice Obrecht, Juan Pablo Giraldo Ospino, Muy-Cheng Peich, Oriana Ponta, Clara van Praag, Kate Radford, Dominic Richardson, Hanadi Riyadh, Beatrice Rutishauser-Ramm, Jasmine Turner, and Nina Weaver

Our acknowledgements would be incomplete without mentioning our team of very able enumerators in Jordan, Kenya, and Rwanda. Specifically, we acknowledge the input of the team of community mobilisers from the World University Service of Canada, teachers from Libraries Without Borders and the Danish Refugee Council, and enumerators from Laterite and Proximity, whose dedication during data collection ensured the collection of high-quality data.

The patience exercised by partner organisations, team contributors, students, and parents during interviews and focus groups is also greatly appreciated. It is our hope that insights garnered from the information they provided will help to enhance the strength of the interventions.

List of Acronyms

AEP	Accelerated Education Programme
AIR	American Institutes for Research
BPRM	Bureau of Population, Refugees and Migration
CWTL	Can't Wait To Learn
DFID	Department for International Development
DRC	Danish Refugee Council
EERCK	Equity in Education in Refugee Camps in Kenya
EGMA	Early Grade Mathematics Assessments
EGRA	Early Grade Reading Assessments
EiE	Education in Emergencies
EoL	Essence of Learning
HEA	Humanitarian Education Accelerator
ICT	Integrated Communication Technology
IRB	Institutional Review Board
KEEP	Kenya Equity in Education Project
KCPE	Kenya Certificate for Primary Education
LWB	Libraries Without Borders
LWF	Lutheran World Federation
MEL	Monitoring, Evaluation, and Learning
MoE	Ministry of Education
M&E	Monitoring and Evaluation
NGO	Non-Governmental Organisation
NN	Nearest Neighbour
NRC	Norwegian Refugee Council
OLS	Ordinary Least Squares
SNHU	Southern New Hampshire University
TNO	Netherlands Organisation for Applied Scientific Research
UN	United Nations
UNEG	United Nations Evaluation Group
UNHCR	United Nations High Commissioner for Refugees
UNICEF	United Nations Children's Fund
WCH	War Child Holland
WDR	World Development Report
WHO	World Health Organisation
WUSC	World University Service of Canada

Contents

Exhibits

Page

Executive Summary

Introduction

In protracted crisis settings, many children and youth lack access to high-quality education: only 63% of all refugee children are enrolled in primary school, compared with 91% of all children globally; only 24% of adolescent refugees attend secondary school; and just 1% of refugees attend university (United Nations High Commissioner for Refugees [UNHCR], 2019). Identifying and scaling effective education innovations could rapidly increase both access to education and the quality of education. Such innovations would be particularly valuable in complex emergencies, as just 4% of the humanitarian aid budget is spent on education (United Nations Children's Fund [UNICEF], 2018). However, major evidence gaps limit our understanding of what works in humanitarian contexts. Only a few studies present credible results on how to scale education innovations in protracted humanitarian crisis settings (de Hoop, Brudevold-Newman, Rothbard, Todd, & Kiggundu, 2018), and there is limited evidence from rigorous experimental or quasi-experimental impact evaluations of education innovations in these contexts (Burde, Guven, Kelcey, Lahmann, & Al-Abbadi, 2015; Elrha, 2018; Puri, Aladysheva, Iversen, Ghorpade, & Brück, 2015). A study by Elrha (2018) also shows that humanitarian organisations currently have insufficient embedded knowledge and skills to support the scaling of innovations.

As a result of these evidence gaps and limited embedded knowledge, only a small minority of promising pilot programmes move to scale (Elrha, 2018; McClure & Gray, 2015a). Although a programme's scale should be relative to the impact that organisations aim to achieve (Elrha, 2018), most of the programmes operating in humanitarian contexts target few students relative to the number of school-aged refugee children who could be served (UNHCR, 2017). Indeed, the number of refugee children reached by humanitarian organisations stands in sharp contrast to the number of school-aged refugee children in humanitarian contexts, as well as the number of school-aged children that many organisations in the scaling space aim to reach. Globally, there are 68 million displaced individuals, including 305,000 school-aged refugee children in Bangladesh; 370,000 school-aged refugee children in Jordan; 240,000 school-aged refugee children in Kenya; 50,000 school-aged refugee children in Rwanda; and over 100,000 school-aged refugee children in Lebanon (UNICEF Bangladesh, 2019; UNHCR Jordan, 2018; UNICEF Kenya, 2019; El Daoi, 2019; UNHCR Rwanda, 2019). Furthermore, the Global Innovation Fund only funds innovations that have the potential to reach millions (Global Innovation Fund, n.d.).

Recognising that there is insufficient knowledge available to inform the scaling of education innovations in humanitarian settings (Elrha, 2018), the Department for International Development (DFID), the United Nations Children's Fund (UNICEF), and the United Nations High Commissioner for Refugees (UNHCR) created the Humanitarian Education Accelerator (HEA), with the aim of understanding how to create the conditions necessary to scale existing pilot programmes. The HEA has two major components:

1. **Research**
    - Providing rigorous evidence of the effectiveness of five selected education interventions and their potential to scale
    - Bringing together findings from HEA research and existing evidence to create a meta-evaluation that examines and summarises barriers and facilitators for scaling education programmes in humanitarian crisis settings
2. **Support and mentorship**
    - Using a mentorship model that pairs subject experts with implementers to mentor, guide, and support them through the scaling process
    - Improving monitoring and evaluation (M&E) capacity through capacity-building and mentorship, and providing M&E funding for selected implementing organisations

This executive summary and the report that follows present the findings of the meta-evaluation conducted under the HEA's research component. The meta-evaluation brings together lessons learned from process evaluations of five implementing agencies ("innovation teams") and impact evaluations of three innovation teams that were selected for the HEA and are in the process of scaling education innovations in humanitarian crisis settings: Caritas Switzerland (Caritas), Kepler, Libraries Without Borders (LWB), War Child Holland (WCH), and the World University Service of Canada (WUSC).

The organisations' education innovations use different strategies to improve learning and psychosocial outcomes for programme participants in a diverse set of contexts. Exhibit 1 depicts the main characteristics of the programmes and their target populations in the contexts studied under the HEA, as well as their outcome measures. For the HEA, we studied the following:

- **WUSC's Equity in Education in Refugee Camps in Kenya (EERCK) programme and Kenya Equity in Education Project (KEEP):** These programmes focus on improving non-cognitive skills (such as aspirations and resilience) and educational outcomes amongst seventh- and eighth-grade girls by providing weekend and holiday remedial education. We studied programme implementation of both programmes and impact of the EERCK programme in Dadaab and Kakuma refugee camps.
- **Kepler's tertiary education programme:** This programme focuses on increasing access to formal employment opportunities by offering a path to a fully accredited bachelor's degree from Southern New Hampshire University (SNHU). We studied programme implementation in Kigali and the Kiziba refugee camp in Rwanda.
- **WCH and partners' Can't Wait to Learn (CWTL) programme:** This programme aims to improve literacy and numeracy outcomes by providing digital, game-based learning via tablets. We studied educational and psychosocial outcomes amongst refugee children and children from the host population residing in Jordan, Lebanon, and Sudan, and we conducted a process evaluation in Uganda.

- **LWB's Ideas Box programme:** This programme focuses on improving psychosocial and learning outcomes for refugee children and children from the host population in more than 20 countries by providing a container-sized box that houses a library and a learning space in a safe environment. We studied programme implementation and impact in Amman, Jordan.
- **Caritas's Essence of Learning (EoL) programme:** This programme focuses on improving childhood development in several countries by stimulating children's non-cognitive and cognitive skills. We studied programme implementation in Bangladesh and Romania.

The innovation team research summaries, which will be published as separate reports, provide more detailed descriptions of the programming and the individual evaluations.

Exhibit 1. Summary of Innovation Designs and Individual Research Designs Under the HEA

Innovation Team (Programme)	Programme Description	Target Group	Number of Programme Participants	Impact Evaluation or Quantitative Study Type	Process Evaluation Type
WCH (CWTL programme)	Digital, game-based learning technology	Out-of-school refugee children in Lebanon, out-of-school children in Sudan, out-of-school refugee children and in-school refugee and host-community children in Jordan, and refugee and host-community children in the Accelerated Education Programme (AEP) in Uganda	2,263 students in Jordan; 6,902 students in Lebanon; 171 students in Sudan; and 3,120 students in Uganda	Quasi-experimental difference-in-difference studies for in-school children in Jordan and out-of-school children in Sudan to estimate impacts on learning and psychosocial outcomes, two proof-of-concept studies in Jordan with in-school and out-of-school children, and a practice-driven evaluation in Lebanon to examine trends in learning and psychosocial outcomes	In-depth case study in Uganda and a study of how lessons learned in Jordan, Lebanon, and Sudan helped with programme implementation in Uganda
WUSC (EERCK programme and KEEP programme)	Weekend and holiday remedial education	Refugee girls enrolled in primary school in Dadaab and Kakuma refugee camps in Kenya	2,560 students annually in KEEP and EERCK; 2,000 new learners every year	Randomised controlled trial in Kakuma and regression discontinuity design in Dadaab to estimate impacts on learning outcomes, school attendance, aspirations, and resilience	Process evaluation to analyse WUSC's implementation model

Innovation Team (Programme)	Programme Description	Target Group	Number of Programme Participants	Impact Evaluation or Quantitative Study Type	Process Evaluation Type
Kepler (Kepler University programme)	Fully accredited bachelor's degree from Southern New Hampshire University	Youth in Rwanda who could benefit from tertiary education	714 non-refugee students and 189 refugee students	Quantitative study to determine employers' knowledge about and attitudes towards recruiting and hiring refugees	Process evaluation to analyse Kepler's journey to scale
LWB (Ideas Box programme)	Portable media centre and learning hub in a safe environment	Syrian, Palestinian, Iraqi, and other refugee children and host-population children in Jordan	327 students in the Marka Johud Centre and 2,614 students in the Danish Refugee Council (DRC) centre	Difference-in-difference analysis to estimate impacts of the self-managed Ideas Box on learning outcomes, and matching combined with difference-in-difference analysis to estimate impacts of the Ideas Box in the DRC community centre on psychosocial outcomes	Process evaluation to understand how different Ideas Box programmes operate within and across different contexts
Caritas (EoL programme)	Holistic pedagogical approach; combines psychosocial services with education	Children in kindergarten and primary school in Bangladesh and Romania	1,892 students in Bangladesh and 431 students in Romania	N/A	Retrospective study to synthesise lessons learned from the experiences of programmes in Bangladesh and Romania

Meta-Evaluation Findings

This report presents the findings from the meta-evaluation (summarised in Exhibits 2 and 3). The results are based on a qualitative synthesis and a quantitative synthesis. The qualitative synthesis used an adaptation of a meta-ethnography approach, based on qualitative data collected as part of process evaluations of the five innovation teams (e.g., Atkins et al., 2008; Brody et al., 2015, 2017). Results were organised into three main themes that influence the scaling process: (1) context, including gender norms and security; (2) business model, including organisational management, financial resources, and partnerships; and (3) programme ownership and advocacy, including community support, demand for the programme, and political buy-in (Ramalingam et al., 2015). The quantitative synthesis estimated the effects of three of the five education innovations on educational and psychosocial outcomes and compared those effects with various systematic reviews of education programmes in low- and middle-income countries (Evans & Popova, 2015; McEwan, 2015; Stone et al., 2018; Snilstveit et al., 2016). We used a randomised controlled trial to estimate effect sizes for WUSC's remedial education programme in Kakuma, quasi-experimental studies to determine the effects of WCH's CWTL programme in Jordan and

Sudan and WUSC's remedial education programme in Dadaab, and proof-of-concept and non-experimental studies focusing on WCH's CWTL programme in Lebanon and LWB's Ideas Box programme in the Danish Refugee Council (DRC) and Johud community centres. We did not conduct a meta-analysis because the innovations focus on a diverse range of topics related to education in refugee contexts (Borenstein, Hedges, Higgins, & Rothstein, 2009; Waddington et al., 2012). We also did not conduct impact evaluations for the Kepler and Caritas innovation teams; claims about the impact of innovations on learning and psychosocial outcomes in this report do not apply to their evaluations.

I. Meta-Evaluation Lessons on Scaling Innovations in Education

Exhibit 2. Key Takeaways on Scaling From the Qualitative Synthesis

- Most innovation teams started multiple pilot projects in different contexts, rather than scaling up in one context. Teams indicated that this approach was adopted, in part, due to strategic considerations, such as the need for codification and the generation of evidence. However, some teams did not have sufficient time and resources to build in key components for scaling their model in one context. As a result, some teams operated in perpetual pilot mode (defined as the implementation of pilot innovations in new contexts, driven by a new emergency or funding opportunity).
- All five innovation teams considered it critical to partner with local and national governments, but the comprehensiveness of these partnerships varied. The nature of the programme and the stage of scaling influenced government engagement strategies. Two teams reported that concerted efforts were required from the outset to maintain relationships with governments. Other teams preferred to resolve organisational and design-related issues and solidify the pilot programme prior to engaging with governments on the potential to scale.
- Innovation teams were flexible and often adapted their programmes based on community demands, and in response to donor priorities.
- Although innovation teams used evidence to various degrees during pivotal implementation moments, the availability of evidence to inform decision-making about education programming remains limited in humanitarian contexts.

Piloting in New Contexts Instead of Scaling Up in the Initial Context

Meta-evaluation findings showed that most innovation teams dedicated time and resources to piloting programmes in new contexts, rather than scaling up programmes in initial contexts. The decision to pilot a programme in a new area (as opposed to reaching more individuals in the current area) appeared to be driven by a combination of strategic considerations, donor location priorities (e.g., grants that support scaling in initial contexts), uncertainty over future funding, and limited access to longer term, flexible funding to establish programme management systems. In some cases, the innovation teams decided to pilot their innovation in multiple contexts to show evidence of effectiveness across multiple settings. This may enable implementers to increase the tangibility or codification of their innovation and reduce complexity, laying the foundations for moving to scale in the initial context (Cooley & Linn, 2014; Results for Development & UNICEF, 2016; Gray, 2019). In other cases, the innovation teams responded to incentives that were driven by the overall aid architecture for education, in which funding is limited and

erratic (McClure & Gray, 2015a; Nicolai & Hine, 2015; Results for Development & UNICEF, 2016) and often provided only for a single pilot programme (McClure & Gray, 2015a). This may increase the risk of operating in perpetual pilot mode, in which pilot innovations are implemented in a variety of new contexts without putting in place the foundations for sustainability.

Adapting programmes to new contexts requires significant time and introduces fixed costs, limiting the potential for economies of scale in some contexts (Tulloch, 2016). The innovation teams spent significant time and resources learning about new contexts, adapting their innovations to those contexts, and building new relationships with government actors. Implementing in new contexts may also require additional research to understand programme effectiveness in the new context (which is costly), recognising that research findings from the initial programme context cannot necessarily be credibly extrapolated to a new context because of limited external validity. Jones (forthcoming) conducted a costing analysis of WCH's CWTL programme, which suggested the potential for economies of scale when scaling up in the initial context, as well as when adapting the same programme to new contexts.

Allocating core funding to establish programme management systems, as well as providing larger, longer term, and unrestricted grants for innovations that are in the final stages of the scaling process, could contribute to an increased likelihood of achieving economies of scale. For example, WUSC was able to focus on scaling up in Kakuma and Dadaab because a large, single, grant-making body prioritised solidifying WUSC's remedial education model in a single location. This grant enabled WUSC to improve programme design before launching pilot projects in new locations.

Evidence-Based Adaptive Programming

Our study found that all the innovation teams had been flexible in adapting their innovations to suit different geographic settings, local needs, and educational priorities. As a result, the five innovation teams were able to secure community support and generate demand for their programmes.

Most adaptations were made in response to community demands or donor priorities. For example, LWB introduced knitting lessons in response to demand amongst women visiting the community centre, and Caritas provided umbrellas to children in response to requests from parents. WUSC shifted the emphasis of its remedial education programme to children with disabilities in response to donor priorities, and WCH chose to implement the CWTL programme in three countries (as opposed to the initially envisioned two countries) in response to donor demands.

Given the limited evidence on the impact of education programmes in humanitarian contexts, using evidence from development settings in low- and middle-income countries and investing in M&E systems can help to strengthen theories of change and encourage evidence-based programming. Investing in the use of evidence and M&E systems also facilitates learning about the emergence of unknown pathways, opportunities, and constraints that can strengthen the initial theory of change. It is also critical to invest in developing business systems that will enable implementers to achieve evidence-based adaptive programming. Ongoing documentation, strategic planning, and financial and other management systems that track programme adaptations can link adaptations to evidence-based theories of change (Vogel, 2012). Currently, most innovation teams adapt and pivot their programmes, but the documentation and strategy for adaptations are not always explicit.

Government Engagement

All five innovation teams considered it critical to partner with local and national governments from the outset, although the comprehensiveness of these partnerships varied. The nature of the programme (e.g., whether it was a formal or non-formal education intervention) and the stage of scaling (pilot versus implementation at scale) influenced government engagement strategies. Two teams reported that concerted efforts were required from the start to maintain relationships with government entities. Other teams preferred to resolve organisational and design-related issues and solidify the pilot programme prior to engaging with governments on the potential to scale.

While teams are gathering evidence and making organisational and design-related adaptations in the pilot phase, they can still increase the likelihood of scaling by aligning programme priorities with public policy priorities and existing national education systems. For example, WCH emphasised the importance of demonstrating its capacity and willingness to adapt to national education systems, which may have contributed to its ability to scale the CWTL programme in Sudan. Postponing close engagement with the government until after a strong pilot programme has been built may limit the implementers' ability to consider how a programme could function in national education systems.

Development of Business Systems

As a result of restricted funding, demand for many of the innovation teams' services outpaced their ability to design and establish systems that could facilitate a smoother scaling journey, such as documentation of organisational, financial, and partnership management; investments in research and development; and other elements of project management that guide programmes as they scale. Building and improving these systems can help to streamline operations, while still leaving room for programmes to adapt. The innovation teams reported a tendency to focus more on outward-facing end products than the underlying systems that would increase the likelihood of high-quality implementation at scale. This meant that almost all innovation teams encountered challenges related to management and implementation throughout programming, which they suggested could have been reduced if funding had been allocated specifically to strengthen organisational capacity in the early stages of their projects.

II. Meta-Evaluation Lessons on the Impact of Innovations in Education

Exhibit 3. Impact Lessons From the Quantitative Synthesis

- Remedial education programmes in Dadaab and Kakuma did not show statistically significant effects on learning outcomes, on average, but combining remedial education with cash transfers or school feeding programmes may create synergies. WUSC's remedial education programme showed positive and statistically significant effects on learning outcomes for girls in food-secure households who attended at least 50 hours of remedial education. It is likely that cash transfers and school feeding programmes would have positive impacts on food security (Gilligan, Margolies, Quinones, & Roy, 2013).
- The effects of technology-in-education programmes on learning outcomes are likely to depend on contextual characteristics, including baseline levels of learning and whether children attend school. CWTL's digital, game-based learning approach showed positive effects on learning outcomes amongst out-of-school children in Sudan. In Jordan, CWTL replaced 40% of classroom time in the

standard government curriculum for public schools, and we found equal learning gains amongst in-school children receiving CWTL and children receiving the standard curriculum. In Lebanon, a before–after study found significant improvements in numeracy and psychosocial outcomes, which key informants attributed to the CWTL programme. However, the study did not include a comparison group.

- Employers in Rwanda with limited knowledge about hiring and recruiting refugees reported that they were less likely to hire refugees than Rwandans with identical characteristics. However, employers who reported having sufficient knowledge about hiring and recruiting refugees were just as likely to hire refugees as Rwandans with identical characteristics. This finding suggests that providing information about hiring and recruiting refugees may contribute to improving refugees' labour market outcomes.

Educational Outcomes

Although we found mixed results regarding impacts on learning outcomes across the different evaluations, the results highlighted several opportunities to increase the likelihood of achieving positive effects. Implementation and contextual challenges may have limited innovation teams' ability to achieve statistically significant effects on learning outcomes. For example, WUSC's remedial education programme in Kakuma and Dadaab refugee camps did not have statistically significant effects on either learning or psychosocial outcomes across the full sample in the refugee camps 1.5 years after the start of the programme. However, the results showed that the programme had positive effects on learning outcomes for girls in food-secure households who attended at least 50 hours of remedial education. For this group, we found statistically significant effects of approximately 0.20 standardised mean differences.[1] As cash transfers and school feeding programmes are likely to produce positive impacts on food security (Gilligan, Margolies, Quinones, & Roy, 2013; Snilstveit et al., 2016), this finding suggests that combining remedial education with cash transfers or school feeding programmes could lead to positive effects on learning outcomes. WUSC has recently started piloting cash transfers as part of its remedial education programme.

WCH's CWTL programme showed comparable learning gains amongst in-school students in Jordan for whom CWTL replaced 40% of classroom time for 3 months and similar students at comparison schools who received the standard government curriculum. This can potentially be explained by four operational challenges identified through the qualitative analysis: (1) limitations of the teacher training, (2) equipment malfunctions, (3) time management difficulties in the classroom, and (4) boredom associated with repetition within the game. Furthermore, impacts were only examined 3 months after the programme began; more research is needed to examine longer term impacts.

WCH's CWTL programme had positive effects on learning outcomes amongst out-of-school children in Sudan. This suggests that multifaceted technology-in-education programmes with a strong focus on pedagogical practices could contribute to improvements in learning outcomes amongst out-of-school students who start from a very low baseline. In a practice-oriented evaluation of WCH's CWTL programme, out-of-school children in Lebanon showed improvements of 0.30 standard deviations in mathematics outcomes 3 months after the start of the programme. Qualitative research attributed these improvements

[1] We exercise caution in interpreting this result, however, because the effects are based on a very small sample (n=170). For this reason, the results were no longer statistically significant when we adjusted them for multiple comparisons.

to the CWTL programme, suggesting that it may have the potential to positively affect mathematics outcomes amongst out-of-school children in Lebanon. We also saw promising trends in Arabic reading and mathematics outcomes amongst out-of-school children participating in LWB's self-managed Ideas Box in the Johud centre in Jordan.[2] Although these positive trends did not allow us to assess the causal effects of the programme, the results suggest that the self-managed Ideas Box may have the potential to positively affect learning outcomes amongst out-of-school children (particularly as the average scores in Arabic and reading outcomes declined in comparison community centres).

Psychosocial Outcomes

The evidence on whether education innovations had positive effects on psychosocial outcomes was mixed. While WUSC's remedial education programme did not have positive effects on resilience, WCH's CWTL programme achieved positive impacts on children's hope in Jordan and positive impacts on children's self-esteem and psychological well-being in Sudan. Qualitative evidence also suggested that the CWTL programme may have contributed to positive trends in psychosocial outcomes in Lebanon. However, results were not consistent across countries: WCH's CWTL programme did not have statistically significant effects on children's self-esteem or psychological well-being in Jordan, nor did it have statistically significant effects on children's hope in Sudan.

We found mixed evidence on the effects on psychosocial outcomes of integrating LWB's Ideas Box programme into the DRC community centre in Jordan. Qualitative evidence attributed improvements in self-discipline and social cohesion to the Ideas Box in the DRC community centre. However, quantitative evidence contradicted the qualitative results, suggesting that the Ideas Box had negative effects on psychosocial outcomes. Methodological limitations of quantitative and qualitative research could explain the differences in findings; more research is required to understand the results. A different study in Colombia found evidence of positive effects on psychosocial outcomes for the Ideas Box (positive feelings about the future, conflict resolution abilities, and participation in the community; CNC, 2018).

Labour Market Opportunities for Refugees

Employers were 7 percentage points less likely to report that they would hire relatively well-educated refugees than nationals with identical characteristics in Rwanda—a context where refugees have the right to work. Analyses suggested that this finding was driven by employers who lacked knowledge about how to recruit refugees in the labour market. Employers who reported sufficient knowledge about hiring and recruiting refugees reported that they were just as likely to hire refugees as Rwandans. This finding suggests that interventions that provide information about the hiring and recruitment of refugees may contribute to improving refugees' labour market outcomes. In contexts with limited legal rights and limited employer knowledge about hiring and recruiting refugees, refugees may face larger challenges in the labour market. Evidence from the Palestinian territories suggests that the economic returns to education decrease dramatically in contexts where refugees do not have the right to work (Angrist, 1995).

[2] LWB is using two implementation models for its Ideas Box programme. The first model emphasises co-creation of Ideas Box programming in close collaboration with non-governmental organisations (NGOs). In the self-managed Ideas Box, LWB creates and implements its own activities.

III. Methodological Research Lessons

Exhibit 4. Methodological Lessons From Qualitative and Quantitative Research

- Lessons learned during the HEA showed that it was particularly important to design flexible evaluations that could adjust to contextual challenges, including unstable operating environments, limited data collection capacity, and frequent implementation delays.
- Collecting data in humanitarian contexts often requires multiple ethical and government approvals, due to the importance of protecting vulnerable populations.
- Experiences during the HEA showed that for data collection purposes, it was important to provide data collection training to—and work with—individuals residing in insecure settings who were trusted by the community and familiar with the context.

We identified three common challenges that can compromise the methodological rigour of evaluations in protracted humanitarian crisis settings, each of which introduced unavoidable delays in the design and implementation of the impact evaluations. First, implementation delays are common in humanitarian crisis settings, in part because it is often challenging to obtain approval to collect data. The number of approvals required for data collection in refugee populations may be even higher than in international development settings because approvals are needed from a larger number of key stakeholders. In Rwanda, for example, data collection approval was required from both the Ministry of Disaster Management and Refugee Affairs and the Ministry of Education (MoE). It was also critical to obtain UNHCR approval to collect data in refugee camps in Kenya, Rwanda, and Jordan.

Second, complex emergencies can create unstable environments that dramatically change planned field activities. For example, disruptive events in the Kiziba refugee camp reduced access to the camp and affected our ability to collect quantitative, household-level data. To adapt in this setting, we primarily relied on qualitative and secondary quantitative data on the Kepler programme.

Third, community distrust of data collectors' motives can make it difficult to collect reliable and valid data in humanitarian contexts. Providing data collection training to people who are trusted and known by communities, and who live within the same setting, can help to address such challenges. Training individuals who reside in refugee camps also limits security concerns and builds data collection capacity in insecure settings.

Conceptual Framework for Scale

We initially adapted a conceptual framework that explained the journey to scale for innovations in education by distinguishing between the design phase (to provide proof of concept of innovations in education), the scaling-up phase, and the scaling-out phase. This initial framework assumed that scaling up would happen before replicating or scaling out programmes in new contexts (see the first column in Exhibit 5). However, most innovation teams started by launching multiple pilot projects in different contexts, instead of scaling up in one context. This decision appeared to be driven by strategic considerations, donor location priorities, and uncertainty over future funding. In some cases, the innovation teams felt it was necessary to demonstrate proof of concept across multiple settings, improve

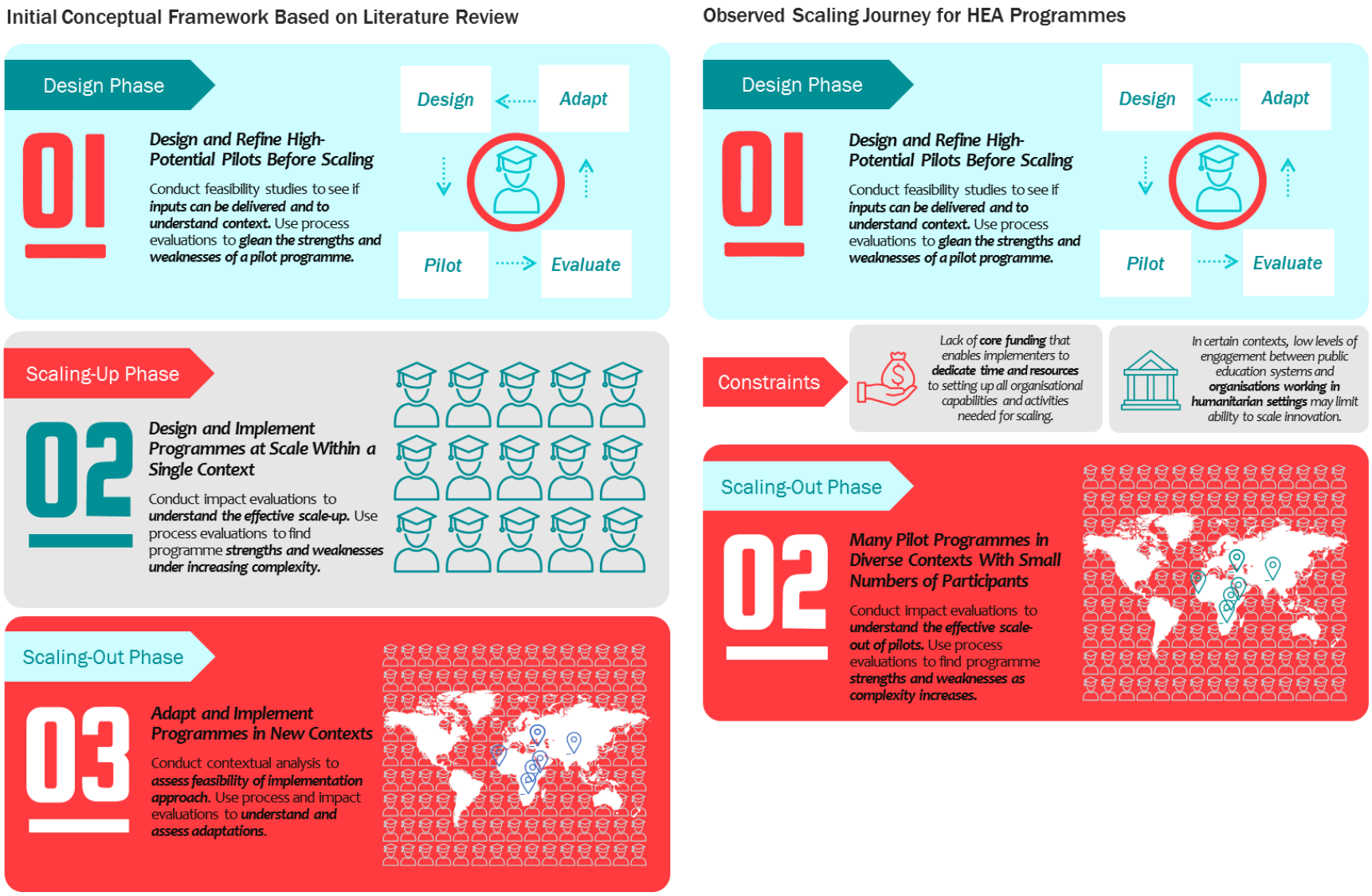
the codification or tangibility of the innovation, and generate evidence to facilitate scaling. Some innovation teams were also incentivised to start multiple pilot programmes in new contexts because of donor priorities and funding uncertainty (reflecting their dependence on grants). In many cases, innovation teams worked simultaneously on scaling up their programme in the initial context and starting new pilots in other contexts, confirming that scaling is not a linear process (McClure & Gray, 2015a; Cooley & Linn, 2014; Results for Development & UNICEF, 2016).

In addition, the innovation teams faced challenges building the management infrastructure and business systems required for scaling innovations in education. These challenges were brought about by limited access to core funding that is not linked to a specific programme but instead enables implementers to dedicate time and resources to setting up all of the organisational capabilities needed for scaling.

We developed a new conceptual framework to reflect these empirical findings (see the second column in Exhibit 5). This new framework highlights that structural incentives reduce the likelihood of expanding education programmes within the same target group, and increase the likelihood of starting multiple pilot programmes in new contexts. It also suggests that it is critical to provide implementers with access to funds to build management systems that enable programmes to scale. To ensure the appropriate use of such funds (i.e., core funding that is not linked to a specific programme but enables implementers to dedicate time and resources to setting up all of the organisational capabilities needed for scaling), implementers can be held accountable for the number of programme participants they reach and improvements in learning and other educational outcomes. In addition, it is critical to continue funding larger scale education innovations in complex emergencies that are close to finalising their scaling journey, while ensuring that enough funding remains available for innovations to show proof of concept in the pilot phase. This new framework also differs from the original framework in its inclusion of structural barriers (such as funding constraints), which limit the ability of innovation teams to scale innovations within the same target group.

The conceptual framework differs from the model of McClure & Gray (2015a), which focused less on where the innovation was used and more on the processes involved during scaling. It also uses a different definition of scaling out: McClure & Gray (2015a) define scaling out as distilling complexity so that the solution is replicable and more easily adoptable, while our framework defines scaling out as the replication of the programme in new contexts.

Exhibit 5. Initial Conceptual Framework Based on a Literature Review Versus the Observed Scaling Journey for HEA Programmes



Recommendations

This report presents seven main recommendations. Results indicate that these recommendations will help implementers to effectively scale education innovations throughout the pilot, scale-up, and scaling-out phases.

1. Donors should allocate core funding that is not linked to a specific programme but instead enables implementers to dedicate time and resources to setting up all of the organisational capabilities and activities needed for scaling, including the development of formalised business systems for improving management, financial, and administrative processes and demonstrating the effects of their innovations on learning outcomes.
2. Donors should continue to provide large-scale funding for education innovations that are further along in the scaling process, and for which there is sufficient evidence of improving learning outcomes in complex emergencies. Donors should also continue to fund smaller scale innovations to reduce the risk of disincentivising early-stage innovation.
3. The international community should encourage and provide pathways to test scaled-up education innovations with governments, and should pilot innovations to help accelerate learning for children and youth in humanitarian settings. Education innovators should start engaging with MoE staff prior to and during implementation to learn about MoE strategy, structure, and priorities, as well as options for sustainability.
4. At the time of its inception, the HEA was one of very few initiatives that aimed to generate rigorous evidence on what works to improve learning outcomes in humanitarian settings. This should remain a top priority amongst education in emergencies (EiE) practitioners and donors, reflected in adequate investment and incentives to generate rigorous evidence.
5. Innovators in humanitarian settings who focus on education should draw on evidence from a wide range of settings—including development settings in low- and middle-income countries—to inform their programme design and strategies and overcome the current evidence gaps in humanitarian contexts.
6. Education innovators in humanitarian contexts should continue to build and strengthen beneficiary feedback mechanisms to improve accountability, focusing not only on communication but also on full disclosure and "bottom-up" dialogue (Heller, Költzow, & Vasudevan, 2011).
7. When starting multiple programmes in new contexts, EiE practitioners should prioritise early development of management systems that are capable of supporting operations in multiple contexts, including streamlining administrative processes, financial tracking, and partnership procedures.

Limitations

This study faced several limitations. First, the research only followed the innovation teams for 3 years, which meant that we could only observe part of their scaling journey. During the research period, many innovation teams moved directly from a pilot phase to a pilot scale-out phase (that is, implementing multiple pilot projects in different countries) without scaling up to reach large numbers of beneficiaries in any one context. As a result, much of the evidence collected stems from the pilot and early stages of the journey to scale, or from the broader literature on barriers to and facilitators of scaling education programmes.

Second, two of the programmes were implemented at limited scale at the beginning of the HEA, and power calculations indicated that we did not have a sufficient sample size to conduct a highly rigorous impact evaluation. As a result, we were only able to design and implement impact evaluations for three of the five innovation teams. Conducting a smaller number of impact evaluations on fewer programmes means that there is less overlap between the different targeted outcomes and limits our ability to conduct a meta-analysis of impact results. In addition, the impact evaluations only covered short periods of time. For WCH, we estimated programme impacts 3 months after the start of the intervention; impact estimates for LWB and WUSC were estimated 6 weeks and 1.5 years after the start of the programmes, respectively. More research is needed to determine the longer term effects of the education innovations.

Third, the small scale of the pilot programmes may limit the external validity of the evaluations. The impact evaluations focused on relatively small-scale programmes that were implemented before scaling up to a larger population. For this reason, it remains unclear whether findings from the impact evaluations can be credibly extrapolated to settings in which the programme is implemented at scale (Banerjee et al., 2017; Bold, Kimenyi, Mwabu, Ng'ang'a, & Sandefur, 2018).

Finally, the programmes included in the study are unlikely to be representative of education innovations in protracted crisis settings. Innovation teams went through a competitive selection process, suggesting that they represent a select group of programmes. For this reason, it is uncertain whether findings apply to other education programmes in humanitarian contexts.

1. Introduction

Many children and youth in protracted humanitarian crisis settings lack access to education. Even when refugee children go to school, providing them with access to quality education is often challenging. In these protracted crisis settings, identifying and scaling effective education innovations could rapidly increase both access to and the quality of education. However, attempts to scale high-potential education innovations in these settings often fail because implementers face complexities that were absent during earlier pilot stages (McClure & Gray, 2015b).

The Humanitarian Education Accelerator (HEA) aimed to generate rigorous evidence on how to effectively scale innovations in education in protracted humanitarian crisis settings. Specifically, the Department for International Development (DFID), the United Nations Children's Fund (UNICEF), and the United Nations High Commissioner for Refugees (UNHCR) established the HEA to identify how to transform high-potential pilot projects into scalable education initiatives for refugees and displaced communities worldwide. To support HEA's goal, the American Institutes for Research (AIR) conducted a meta-evaluation, drawing lessons from process and impact evaluations of five implementing agencies ("innovation teams") that were selected for the HEA and are in the process of scaling high-potential innovations in humanitarian crisis settings: Caritas Switzerland (Caritas), Kepler, Libraries Without Borders (LWB), War Child Holland (WCH), and the World University Service of Canada (WUSC).

Common themes assessed across innovation teams included the journey to scale, business models, ownership and advocacy, and the innovation's impacts on educational and psychosocial outcomes. The teams and their associated innovations are summarised in Exhibit 6. More details on each of the innovation teams are presented in team summaries, which will be published separately. Exhibit A-1 in Annex A presents an overview of AIR's and the innovation teams' respective roles in the evaluation.

This report presents meta-evaluation evidence based on quantitative and qualitative data collected as part of the process evaluations of the five innovation teams, as well as a randomised controlled trial, three quasi-experimental studies, a non-experimental study, and two costing analyses. We present process evaluation findings related to three main themes that affect the scaling process: (1) context, including legal and institutional structure and security; (2) business model, including organisational management, financial resources, and partnerships; and (3) programme ownership and advocacy, including community support, demand for the programme, and political buy-in. We also present impact evaluation findings that focus on educational and psychosocial outcomes for three of the five innovation teams: WCH, WUSC, and LWB. We did not conduct impact evaluations for the Kepler and Caritas innovation teams; any claims

about the impact of innovations on learning and psychosocial outcomes in this report do not apply to their evaluations.

Exhibit 6. Summary of the Innovation Designs and Evaluations

Innovation Team (Programme)	Programme Description	Target Group for Evaluation	Impact Evaluation or Quantitative Study Type	Process Evaluation Type
WCH (CWTL programme)	Digital, game-based learning technology	Out-of-school refugee children in Lebanon, out-of-school children in Sudan, out-of-school refugee children and in-school refugee and host-community children in Jordan, and refugee and host-community children in the Accelerated Education Programme (AEP) in Uganda	Quasi-experimental difference-in-difference studies for in-school children in Jordan and out-of-school children in Sudan, two proof-of-concept studies in Jordan with in-school and out-of-school children, and a practice-driven evaluation in Lebanon with out-of-school children	In-depth case study in Uganda and a study of how lessons learned in Jordan, Lebanon, and Sudan helped to scale out the programme to Uganda
WUSC (EERCK programme and KEEP)	Weekend and holiday remedial education	Refugee girls enrolled in primary school in Dadaab and Kakuma refugee camps in Kenya	Randomised controlled trial in Kakuma and regression discontinuity design in Dadaab	Process evaluation to analyse WUSC's implementation model
Kepler (Kepler University programme)	Fully accredited bachelor's degree from Southern New Hampshire University	Youth in Rwanda who could benefit from tertiary education	Quantitative study to determine employers' knowledge about and attitudes towards recruiting and hiring refugees	Process evaluation to analyse Kepler's journey to scale
LWB (Ideas Box programme)	Portable media centre and learning hub in a safe environment	Syrian, Palestinian, Iraqi, and other refugee children and host-population children in Jordan	Difference-in-difference analysis for the self-managed Ideas Box, and matching combined with difference-in-difference analysis for the Ideas Box in the Danish Refugee Council (DRC) community centre	Process evaluation to understand how different Ideas Box programmes operate within and across different contexts
Caritas (EoL programme)	Holistic pedagogical approach; combines psychosocial services with education	Children in kindergarten and primary school in Bangladesh and Romania	N/A	Retrospective study to synthesise lessons learned from the experiences of programmes in Bangladesh and Romania

Our findings represent a substantial increase in the evidence base on how to scale education innovations in protracted humanitarian crisis settings. Research at the start of the HEA identified only nine studies that presented high-quality evidence on the barriers to and facilitators of scaling in such settings (de Hoop et al., 2018). Tracking multiple innovation teams over several years has enabled us to identify a number of

new barriers and facilitators, and to better understand how the barriers and facilitators identified in earlier studies interact during the scaling process.

The meta-evaluation showed that most of the innovation teams started new pilot programmes in multiple contexts, instead of scaling up in one context. It was common for teams to implement new pilot programmes in multiple settings either because of strategic considerations or because the innovation team received many small grants or inflexible funding from donors with specific location priorities, providing incentives to implement programmes in specific contexts. The teams adapted their programmes to those specific contexts and engaged with governments using a diverse set of strategies. However, the combination of strategic considerations, donor incentives, and uncertainty over future funding limited their ability to scale up to serve a large number of people in their initial contexts, increasing the risk of operating in perpetual pilot mode (Gray, 2019). Importantly, some teams reported that they strategically chose to pilot across multiple settings to increase the tangibility or codification of their innovation in multiple contexts, reduce complexity, and facilitate the scaling of their innovation. This strategy may be beneficial in the initial stages of the scaling process, because scaling may take time. Nonetheless, it is important to lay the foundations for a sustainable model early by engaging closely with governments, collecting evidence on the impact of the programme, and setting up formalised business systems (facilitated through longer term, flexible funding) when piloting across multiple contexts.

The remainder of this report is structured as follows. We begin by presenting our conceptual scaling framework, followed by a discussion of the meta-evaluation findings, based on a synthesis of quantitative and qualitative studies. The report concludes with a discussion and recommendations.

2. Conceptual Framework

Existing Conceptual Frameworks for the Effective Scale-Up of Education Programmes

An evidence synthesis on how to effectively scale education programmes in humanitarian contexts must be built on a conceptual framework that includes the pathways to scale-up, the barriers to and facilitators of scaling education innovations, and the iterative steps required. In developing the conceptual framework, we relied on a combination of the conceptual frameworks of McClure and Gray (2015b) and Pritchett, Samji, and Hammer (2012), as well as a literature review on how to effectively scale education innovations in humanitarian contexts (de Hoop et al., 2018). McClure and Gray (2015b) focus on the complexities of scaling programmes through an innovation lens, while Pritchett and colleagues (2012) focus on an iterative monitoring and evaluation (M&E) framework that builds on the development economics literature. We supplemented these frameworks with studies on scaling by Ramalingam and colleagues (2015) and Obrecht, Warner, and Dillon (2017), who focus on the scaling of innovations. In February 2017, we identified further barriers and facilitators through an evidence synthesis on how to effectively scale innovations in education in low- and middle-income countries (de Hoop et al., 2018). We updated this evidence synthesis in April 2019. In the text box that follows, we present a summary of the results of the initial evidence synthesis, which focused on evidence outside the HEA. Annex H summarises the methodology used for the initial and updated evidence synthesis.

Summary of the Literature Review on How to Scale Innovations in Education

Design Phase

- Insecurity can limit or preclude participation in education and labour market programmes.
- High mobility can limit the effectiveness of innovations in education, because it can prevent students from finishing education and limit students' ability to obtain access to education in their own language.
- Failing to account for existing gender norms can reduce programme participation, increase dropout rates, and limit desired behaviour change.

Scaling-Up and Scaling-Out Phases

- Different operating models and constraints on government employees could hinder the integration of promising pilot programmes into national education systems, if pilot programmes do not adapt.
- Refugees often encounter major challenges when they wish to enter education or the labour market because they have limited rights.
- A lack of involvement amongst community members, local institutions, and stewards of community resources can limit buy-in and ownership.
- A variety of programmes struggle because of limited human or material resources, inconsistent programme funding, high costs, and a potential lack of sustainability.
- Programme flexibility can facilitate scaling, particularly when combined with close engagement with the government, and informed by evidence from evaluations.

Conceptual Framework of the HEA

The Three Phases of Scaling: Design, Scale-Up, and Scale-Out

To define the different phases in the scaling process, we distinguished between the design phase, the scaling-up phase, and the scaling-out phase. The design phase focuses on the design and implementation of a pilot programme (Pritchett et al., 2012), optimising the processes that will be used in the programme's implementation and providing proof of concept. The scaling-up phase focuses on expanding the programme within the same target group. The scaling-out phase focuses on adapting the programme to enable effective implementation in a different context. McClure and Gray (2015b, p. 9) describe scaling up and scaling out as the "missing middle" of the scaling process and note that successful scale-up of education programmes during these phases often requires programme adaptations to account for the increased complexity of working with more participants.

Our conceptual framework differs from McClure and Gray's (2015a) model, which focused more on the processes involved during the scaling process, rather than where the innovation was used. McClure and Gray (2015a) also define scaling out as distilling complexity so that the solution is replicable and more easily adoptable, while we define it as the replication of the programme in new contexts.

Design Phase: Decisions made during the design phase can have significant implications for the success of the pilot and the later scaling phases. For example, ensuring that the intervention is contextually appropriate can improve programme effectiveness; learning by identifying and overcoming implementation constraints can inform subsequent scaling processes; and involving governments in pilot design can help to secure official involvement and buy-in for future scaling plans. During the design phase, it is also crucial to develop a theory of change that hypothesises how the programme will achieve its intended objectives by mapping out the causal chain of inputs, activities, outputs, outcomes, and impacts, as well as the underlying assumptions (White, 2009). Developing the theory of change requires implementers to think backwards to identify which assumptions need to be fulfilled in order for the programme to achieve its intended outputs, as well as its intermediate and final outcomes (Pritchett et al., 2012).

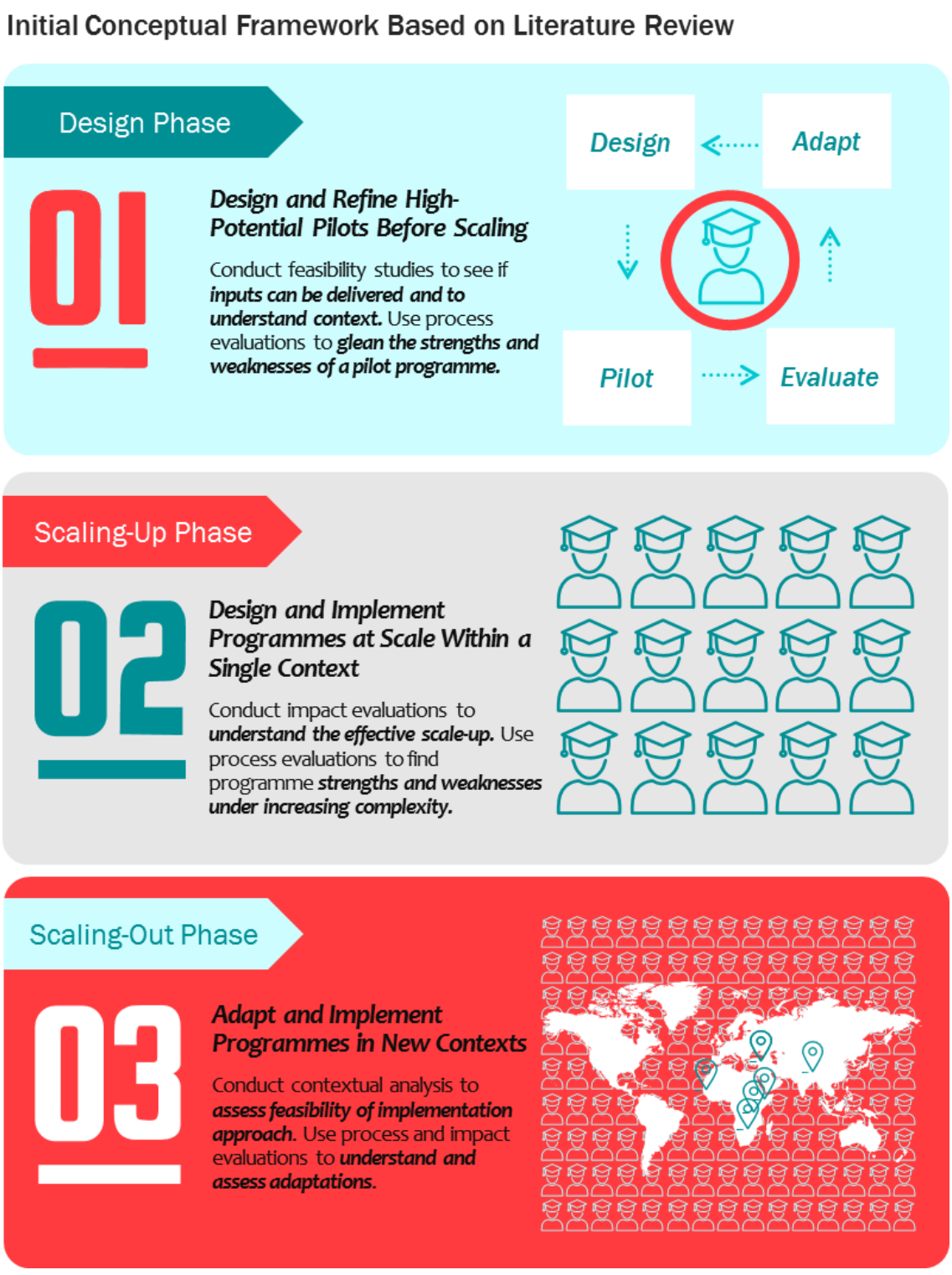
Scaling-Up Phase: The scaling-up phase focuses on adding complexity to create a complete and sustainable solution in order to expand a programme within the same context, adapt a programme to suit a different target group, or collaborate with a different implementing agency, such as the government. In each of these cases, the additional participants and increased number of stakeholders—as well as the need to build a complete solution—add complexity to the implementation, relative to a well-functioning pilot. Both internal and external factors increasingly influence the implementation and effectiveness of a programme during this scale-up phase (McClure & Gray, 2015b). For example, moderating and mediating factors may change when a programme is transferred to other implementing agencies. In Kenya, a contract teacher programme showed fewer benefits when it was implemented by the government, compared to when it was implemented by non-governmental organisations (NGOs), because of challenges associated with integrating the programme into national education systems (which required formalisation). In the scaling-up phase, union membership emerged as an important moderating factor—one that did not have an impact during the pilot phase because of the less formal nature of the programme. Conversely, teacher motivation—an important mediating factor during the pilot phase—was no longer positively affected when the programme was implemented at scale (Bold et al., 2018). In contrast, Kenya's Tusome national literacy programme was able to scale up successfully by galvanising the support of key actors in the education sector to promote a small number of evidence-based interventions intended to improve teacher pedagogy (Piper, Destefano, Kinyanjui, & Ong'ele, 2018). Including training and classroom support (delivered by government officers) from the outset enabled the team to identify cost-effective programme elements that were feasible to implement through national education systems (Piper et al., 2018).

Scaling-Out Phase: In the scaling-out phase, some of the increasing complexity (from internal and external factors) can be distilled through adaptations to implementation that allow the programme to be replicated more easily in different contexts (McClure & Gray, 2015b).[3] This includes distilling the complexity inherent within the solution itself. As programmes can be replicated in this final phase, further programme adaptations may be required to suit different contexts. For example, successful implementation of BRAC's graduation approach—a programme that was developed in Bangladesh and provided asset grants, training and support, life skills coaching, temporary cash consumption support, and access to saving accounts—in other locations required several programme adaptations (Banerjee et al., 2015).

We summarise the conceptual framework in Exhibit 7. The framework is underpinned by the idea that the first step in the scaling process is to define the programme's big-picture goal and then identify the pathways in the theory of change that contribute to achieving that goal. These pathways may change, depending on the phase of the scaling process.

[3] "Context" does not necessarily refer to geographical context. For example, "scaling out" can also refer to the implementation of a programme in an informal education context, as opposed to a formal education context.

Exhibit 7. Initial Conceptual Framework Based on a Literature Review



Evaluation and Adaptation in the Scaling Framework

Each phase requires different types of evaluation and a willingness to adapt the programme as needed. Early in the design phase, feasibility studies or needs assessments can help to inform decisions about whether and how a programme should be implemented. These studies guide implementers in their assessment of whether it is possible to implement an intervention effectively in a particular context, and in their refinement of the theory of change.

When implementing a pilot, it is important to critically examine and document implementation. This can be done though process evaluations, which establish whether the intervention can be improved, identify any new challenges that have emerged, and determine the extent to which the intervention reaches its target population. During this phase, programme adaptation may again be necessary, as well as further changes to programme design and a new feasibility study if programme adaptations are substantial. For these adaptations, it is important to set up organisational management systems to learn from failure and facilitate the journey to scale for promising development programmes in protracted humanitarian crisis settings (Bessant et al., 2014; Ramalingam et al., 2015). Organisational management systems involve collaboration with others, the organisation of an innovation process, the generation and use of evidence, and engagement with end users (Obrecht & Warner, 2016).

The conceptual framework suggests that both impact and process evaluations are useful learning tools during the scale-up phase, when hidden complexities and new challenges may emerge. For example, scaling up may require additional partners, which may change how a programme is implemented; changes in programme implementation may in turn require refinements to the theory of change. Conducting an impact evaluation during this phase is important to examine the programme's pathways and determine whether it achieves its intended outcomes. Conducting a process evaluation during this phase is important because careful re-examination of pathways in the theory of change is necessary, and because a process evaluation can examine any new strengths, weaknesses, and challenges brought about by implementation on a larger scale.

3. Methods

Qualitative Methods

The qualitative synthesis organised findings by theme based on the evidence synthesis (de Hoop et al., 2018). We examined barriers to and facilitators of scaling across three domains, based on studies by Ramalingam and colleagues (2015) and Obrecht and associates (2017): (1) context, (2) business model, and (3) advocacy and ownership. The synthesis used a sequential data collection strategy, which enabled more in-depth investigation and identification of lessons learned regarding the barriers and facilitators identified in previous rounds of data collection.

The three domains were selected to encompass the barriers to and facilitators of scaling innovations identified in the initial evidence synthesis (de Hoop et al., 2018). The evidence synthesis noted the influence of the following factors on the scaling process: gender norms (Chaffin, 2016; IDinsight, 2016; Chinen, Coombes, de Hoop, & Elmeski, 2016; Crea & McFarland, 2015; Pick de Weiss, Andrade-Palos, Townsend, & Givaudan, 1994; Dupas, 2011); security concerns (Chaffin, 2016; American Institutes for Research, 2016a); mobility (Crea & McFarland, 2015); refugee legal rights (IDinsight, 2016; Purnell & Kengkunchorn, 2008); integration into national education systems (Bold et al., 2018; Cameron & Shah, 2017; Humphreys, Sanchez de la Sierra, & Van der Windt, 2014; Lacey, Cooper, & Torrance, 1993); community support (Li, 2012; Agha, 2016); limited implementation capacity (Stubbé, Badri, Telford, Van der Hulst, & Van Joolingen, 2016; Grootenhuis & Calo, 2016; MacLaren, 2010); limited or inconsistent programme funding and infrastructure (American Institutes for Research, 2016b; Refugee Studies Centre, 2005); and flexibility and government buy-in (Banerjee et al., 2017; Chau, Seck, Chandra-Mouli, & Svanemyr, 2016; Epstein, 2014; Gove, Poole, & Piper, 2017; MacLaren, 2010; Piper & Mugenda, 2013; Winthrop & Kirk, 2008).

Factors identified in the literature as influential for the scaling process:

- Gender norms
- Security concerns
- Mobility
- Refugee legal rights
- Integration into national education systems
- Community support
- Implementation capacity
- Programme funding and infrastructure
- Government buy-in

The domains and factors for scaling enabled the research team to create a common coding structure (see Annex F). In addition to the factors identified in the evidence synthesis, we included the following factors

as hypotheses, based on our knowledge of the innovation teams: technology, cultural norms, social exclusion, and the future prospects of participants were included in the context domain; demand was included in the advocacy and ownership domain; and exit strategy and project design and implementation were included in the business model domain. These nodes continued to evolve throughout the coding and analysis process, based on evidence generated from the innovation teams.

The results section in this report includes tables that indicate the current strength of the evidence for each type of barrier or facilitator. We label evidence as weak (indicated by a red circle) if only one of the innovation teams highlighted a barrier or facilitator, moderate (indicated by a yellow circle) if two or three of the innovation teams highlighted a barrier or facilitator, and strong (indicated by a green circle) if four or all of the innovation teams highlighted a barrier or facilitator. This approach aligns with an adaptation of the meta-ethnography approach used by Atkins and colleagues (2008) and Brody and colleagues (2015, 2017).

Quantitative Methods

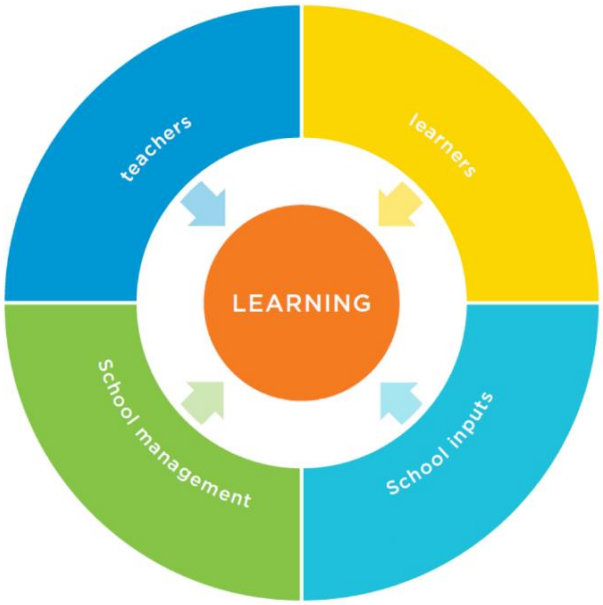
The quantitative synthesis explored the quantitative effects of the innovations using a narrative synthesis approach, estimating effect sizes for each of the outcome measures and comparing the estimated effects with those of comparable programmes. This comparison enabled us to assess the external validity of the impact estimates. We did not conduct a meta-analysis because the innovations focus on a diverse range of topics related to education in refugee contexts (Borenstein et al., 2009; Waddington et al., 2012). Instead, we synthesised the global literature on the effects of education programmes in low- and middle-income countries with results from three of the innovation teams: WUSC's EERCK programme, LWB's Ideas Box programme, and WCH's CWTL programme. The synthesis included evidence on the effects of the EERCK programme in Kakuma and Dadaab; the effects of the CWTL programme in Jordan, Lebanon, and Sudan; and the effects of the LWB-managed Ideas Box and the DRC-managed Ideas Box in Amman, Jordan. We also present evidence on the labour market prospects of refugees in Rwanda. In some cases, we incorporated research designed and implemented through multi-partner research teams (including the innovation teams) in the quantitative synthesis.

In addition, we include the results of costing and cost-effectiveness analyses, in which we estimated the costs of replicating the Kepler programme outside Kiziba refugee camp, and Jones (forthcoming) assessed the cost-effectiveness of WCH's CWTL programme. We also include an analysis by Jones (forthcoming) examining the ability of the CWTL programme to achieve economies of scale.

We examined areas for improvement or potential synergies with other education or multi-sectoral programmes by identifying partnerships that could increase the effectiveness of the innovation teams' programmes. This approach aligns with findings from the latest World Development Report (WDR) on education, which highlights how educational outcomes are directly affected by the quality of school inputs, school management, and teachers, as well as the education preparedness of learners (World Bank, 2018).

In theory, improvements in the quality of any one of these factors could lead to improvements in learning outcomes (Exhibit 8; World Bank, 2018). However, the WDR presents evidence showing that improvements in learning outcomes are likely to require improvements across each factor (World Bank, 2018). This is consistent with the findings of Snilstveit and colleagues (2016), who note that education programmes are unlikely to improve learning outcomes unless they ease multiple constraints. For this reason, we linked the quantitative evidence to the World Bank's conceptual framework by interpreting quantitative findings in light of the implementation challenges identified in the qualitative research. We also linked the results of an employer survey in Rwanda to labour market opportunities for refugees in humanitarian settings where they are allowed to work.

Exhibit 8. Conceptual Model for Improving Learning Outcomes



Note. Reprinted from World Bank (2018).

4. Results

Qualitative Results

I. Programme Context and Scale

Research Questions

- How have contextual factors influenced programme implementation for innovation teams?
- What lessons learned about these contextual factors can facilitate the scale-up of HEA programmes?

Contextual factors such as gender norms, security, refugee legal rights, and infrastructure can influence the successful scaling of promising interventions in humanitarian contexts (de Hoop et al., 2017; McClure & Gray, 2015c). Our analysis identified the following overarching themes related to contextual factors that affected programme implementation:

- Teams prioritised the adaptation of programmes to align with or complement the national education curriculum, depending on government requirements and the availability of public schooling for target populations. In some cases, however, innovation teams struggled to match content to students' learning levels and language abilities. Additionally, some teachers in formal schools found it difficult to integrate the interventions into their existing teaching timetables.
- Local staff and partners brokered access to sites in insecure settings and shared or co-financed space and limited resources such as facilities, electricity, and Internet access. Partners also led the push for relevant content and managerial adaptations to programmes in new contexts. Innovation teams increasingly looked to local partners to take over local operations and financing as a mechanism to sustainably scale their programmes.

This section discusses common contextual challenges, facilitators, and subsequent adaptations during implementation and scaling related to (1) legal and institutional structures, (2) security and physical resources, and (3) gender norms.

a. Legal and Institutional Structures

Existing evidence suggests that restricted access to educational and employment opportunities creates challenges that programmes must address in their design or adapt to when scaling (IDinsight, 2016; Purnell

& Kengkunchorn, 2008). When host countries limit the right to work, the benefits of education become increasingly unclear, and demand for—and the effects of—education or training programmes may weaken (Angrist, 1995). In addition, employers' limited knowledge of how to recruit and hire refugees can limit refugees' ability to participate in the labour market, even in contexts without legal restrictions.

Innovation teams worked during the pilot stage in each location to maximise the likelihood that refugees could participate in education and labour markets in host countries. This section outlines their approaches.

i. Adapting Programmes to National Education Systems

Innovation teams adapted their programmes to align with or complement national curricula in countries of operation.[4] Existing literature demonstrates the need to adapt pilot activities to inform scale-up and promote government buy-in (Gove et al., 2017; Tinajero, 2010). When scaling out the CWTL programme, WCH adapted literacy and mathematics games to align educational content with national curricula. This helped WCH secure support from ministries of education (MoEs), parents, and teachers, and positioned the programme well for eventual integration into national education systems. Similarly, remedial classes for the EERCK and KEEP programmes[5] in Kenya followed the local curriculum to prepare students for national examinations.

LWB and Caritas aligned their approaches to complement government curricula, recognising that complete alignment was neither required nor desirable. In Jordan, LWB designed activities based on the MoE-approved non-formal curriculum for the self-implemented Ideas Box in a community centre that provided non-formal education. Teachers held classes in the Ideas Box once a week and worked with facilitators to use the content in the box to enhance learning. Caritas's EoL curriculum did not mirror the official education curriculum in Romania because it was primarily a social support programme, rather than an education programme. However, Caritas worked with the government to demonstrate how the programme complemented the national education system and enabled Roma students to perform better in public schools.

Evidence from the HEA adds to the existing literature (Bold et al., 2018; Banerjee et al., 2017) by identifying challenges associated with integrating programmes into national education systems. For example, WCH and partners integrated the CWTL programme into the existing Accelerated Education Programme (AEP) curriculum in Uganda and the formal education system in Jordan so that the programme could replace instruction in several sessions per week. While the content was carefully aligned, teachers in both countries had difficulty understanding that CWTL replaced curriculum content and therefore still felt compelled to teach according to the set lesson plan.

Programme content was offered in languages that would enable mainstreaming into national systems. Two innovation teams (WCH and WUSC) delivered programmes to learners in the host country's language of instruction to align with MoE practice and help to mainstream students into national systems. CWTL staff in Uganda said the youngest students tended to have limited exposure to English before enrolling in AEP classes and experienced difficulties learning in English: "Some learners, especially for level 1, may not really understand or really listen to English because they don't know English when they are coming in." In

[4] The Kepler programme offers access to a US-accredited degree from SNHU and does not necessarily align with Rwandan degree programmes.

[5] EERCK targets medium-performing girls; KEEP targets low- and medium-performing girls.

response, teams provided ad hoc language support for students to bridge the gap between the language of instruction and refugees' mother tongue. Several studies note the benefits of supporting mother tongue learning (Benson, 2009; Ouane & Glanz, 2010; Nakamura, de Hoop, & Udayakumar Holla, 2018). As students in Kenya and Uganda tended to be unfamiliar with the host country's language of instruction, WUSC and CWTL implementing partners tried to place refugee teachers in classrooms to explain concepts in students' mother tongue. A remedial teacher in Dadaab explained: "We talk to them in English when we are teaching in English, but when it comes to subject like Kiswahili, we do translate that to their mother tongue for them to get better."

Programmes expanded to benefit both refugees and host-community students, demonstrating the potential to reach a large number of participants as programmes scale up and bridge the humanitarian–development nexus. During the scaling process, some innovation teams broadened their scope to include more students from host communities in certain contexts (WUSC and WCH) or shifted from serving primarily host communities to also including refugees in the same context (Kepler and LWB). WUSC implemented the remedial KEEP programme in refugee camps, as well as in neighbouring host communities. The emphasis on education for host-community students led to strong relationships with the local government and the MoE. Staff noted that providing teachers and undertaking construction in host schools facilitated government buy-in:

> *The issue of staffing in the school was a serious issue when we were coming in. Schools don't have teachers because of the insecurity issue in the region. And we constructed some classrooms in the schools, the latrines, toilets for the girls and other things.*

In Uganda, primary schools in refugee settlements served mostly refugee children. However, some host-community students enrolled in the same schools because of the Refugee and Host Population Empowerment framework, which mandates that 30% of all refugee interventions benefit the host community.

In other cases, teams integrated refugees and host-community students and noted positive interactions between participants. Three innovation teams (LWB, Kepler, and WCH) served refugee and host-community students together in some programme settings, striving to serve students with the greatest need regardless of background. In Jordan, Ideas Box facilitators noticed improved social cohesion amongst Jordanians and Syrians who worked together in the Ideas Box during school holidays. One refugee student from Kepler Kigali noted the benefits of being exposed to Rwandan students in the classroom:

> *We learn how to relate with other students from different backgrounds who are not refugees because we live together in our rooms and share meals. There are many activities we do at the beginning of the term that make students work in groups and get to know each other.*

Communication and sensitisations helped to mitigate feelings of exclusion amongst non-participants and facilitate community buy-in.

ii. Employment Prospects

Programmes added skill-related components to increase the likelihood of contributing to positive livelihood outcomes for refugees. Kepler and LWB provided skill-building opportunities to prepare participants for the job market. In the Ideas Box in Jordan, LWB incorporated Edraak—an open online course platform that provides an employer-recognised certificate. One respondent said that beneficiaries were interested in using Edraak to improve their career development skills. Similarly, Kepler Kiziba students considered themselves more prepared for the job market because of the programme's emphasis on workforce-relevant skills and Kepler's arrangement with employers to hire refugee interns. A study by IDinsight (2019) also showed that SNHU–Kepler graduates performed better on academic assessments, such as English reading and writing, mathematics, computer literacy, and critical thinking, and had better employment prospects than a matched comparison group of students enrolled in other universities in Rwanda.

In countries where refugees lack the right to work, innovation teams motivated students to pursue education through scholarship opportunities or connected them to work opportunities within the camp itself. WUSC provided scholarships to attend university in Canada to high academic achievers in secondary schools in Kakuma and Dadaab refugee camps. The scholarship opportunity encouraged girls enrolled in WUSC's remedial education programmes (EERCK and KEEP) to complete primary and secondary school in order to qualify for the scholarship. Several primary-school-aged girls who dropped out of the KEEP programme said that they were motivated to return to remedial classes by the WUSC scholarship: "If I get a chance, I will go back to the remedial classes because now [I] am in class 8. If I get good marks, I can also get [a] scholarship. That's why I encourage myself to go there."

Exhibit 9 summarises the findings related to legal and institutional structures.

Exhibit 9. Legal and Institutional Structures: Findings and Strength of Evidence

Finding	Innovation Teams	Strength of Evidence From HEA
Incorporating MoE-approved curricula into education programmes was often a pre-condition for scaling.	• WCH • WUSC • LWB • Caritas	
Programme content was offered in languages that would enable mainstreaming into national systems.	• WCH • WUSC • Kepler	
Programmes expanded to benefit both refugees and host-community students, demonstrating the potential to reach a large number of participants as programmes scale up.	• WCH • WUSC • LWB • Kepler	
Teams increased communication and sensitisations to reduce feelings of exclusion amongst non-participants and facilitate community buy-in.	• WUSC • LWB	
Programmes added skill-related components to increase the likelihood of contributing to positive livelihood outcomes for refugees.	• Kepler • WUSC • LWB	

b. Security and Physical Resources

Insecurity can reduce programme participation if participants lack a safe method of getting to and from school, and can decrease the willingness of skilled staff to work in programme areas (Chaffin, 2016; Bennell, 2004; Ring & West, 2015). Programmes attempting to scale in refugee and protracted crisis settings must also undertake implementation with limited resources, including facilities, electricity, and Internet access (Pingel, 2010; Dahya, 2016). Efforts to mitigate these challenges help to maintain programme quality when scaling.

Partnerships facilitated access to insecure locations, resources, and facilities during pilot and design phases. Two teams reported that identifying partners with a deep understanding of the local context and security situation supported scaling. For example, LWB discussed how it worked with local liaisons to gain access to implementation sites when scaling out. Similarly, a respondent from WUSC described the benefits of employing local staff, especially during times of insecurity: "Like most of the remedial teachers in Dadaab, they are locals. So, when there is insecurity, they are not necessarily affected by that [and can continue to run classes]."

All teams depended on contributions from partners or worked with partners to co-finance expenses for limited resources, including electricity and facilities. Under this model, teams determined the best way to use existing resources during the design phase. Kepler overcame challenges with electricity by borrowing a generator from another organisation in Kiziba camp. However, our costing analysis showed that installing solar equipment at the beginning of the programme could have saved Kepler's operation budget more than $10,000, relative to relying on the generator at start-up. Teams also implemented their programmes in spaces identified or provided by local partners. Two innovation teams used existing school facilities through agreements with partners: remedial classes for KEEP and EERCK were hosted in primary schools, while CWTL was implemented in AEP centres in Uganda and a range of NGO-led centres in Lebanon. Similarly, partners typically provided a space in which to house the Ideas Box and run activities.

Wraparound services encouraged student participation. Three teams (WUSC, Kepler, and Caritas) included a food security component and distributed learning materials to students. The provision of food aligns with existing evidence that school feeding programmes increase enrolment, attendance, and learning outcomes in low- and middle-income countries, as well as in humanitarian contexts (Snilstveit et al., 2016; Tull, 2018). Two teams provided more nutritious options or meals for students, while another team provided only a snack (with less nutritional value). Kepler provided students with laptops, course materials, and access to services such as supplemental healthcare and counselling. Teachers, community mobilisers, community leaders, girls, and staff agreed that the distribution of school supplies, sanitary pads, and book lights helped to motivate girls to attend remedial classes under the KEEP programme.

Exhibit 10 summarises the findings related to security and physical resources.

Exhibit 10. Security and Physical Resources: Findings and Strength of Evidence

Finding	Innovation Teams	Strength of Evidence From HEA
Partnerships facilitated access to insecure or hard-to-reach locations, resources, and facilities during the pilot and design phases.	• WCH • LWB • Kepler • Caritas	
During the scaling process, teams provided in-kind support to facilitate student participation.	• WUSC • Kepler • Caritas	

c. Gender Norms

Gender norms that assume women are primarily responsible for activities such as cooking, cleaning, childbearing, child-rearing, and caring for family members can impede female participation in education programmes during all phases of the scaling process (Chaffin, 2016; Idinsight, 2016). Kepler and WUSC initially had programme components that exclusively targeted girls, but these programmes caused non-participants to feel excluded. Kepler opened access to all students for its preparatory programme but continued to have increased and specialised programme time reserved for female students. WUSC increased communication with non-participants to explain how the specific challenges faced by girls warrant additional support compared to boys. Programme staff and community members from WUSC said opposition from non-participants fell after attempts were made to counteract perceptions of exclusion through additional sensitisations with non-participants.

Sensitisations targeting parents and guardians increased support for girls' education. Respondents from Kepler and WUSC reported that the burden of household responsibilities often affected attendance and academic performance amongst female students. WUSC relied on community mobilisers to speak directly to parents/guardians when girls did not attend classes regularly or dropped out, and to spread awareness about the benefits of girls' education in the community. Community mobilisers believed that girls were less likely to miss classes after these sensitisation efforts because parents and guardians placed a higher value on girls' education.

Specialised support decreased dropout amongst women and girls and reduced barriers to scaling. Three innovation teams (WUSC, Kepler, and Caritas) experienced some dropout amongst women and girls due to early marriage and pregnancy. A respondent from WUSC explained how project staff encouraged young mothers to continue their education after childbirth:

> *We do follow-ups and encourage [girls who have given birth] to get back to school … And there are times when they [girls] need to leave school earlier to go and take care of the babies or the younger siblings, so we do allow them. So we have that kind of flexibility, which also gives them assurance that [they] can go to school.*

The Kepler programme now includes maternity leave for young mothers to encourage them to complete their degrees and internships—an example of a useful adaptation during scaling.

WUSC also started to distribute sanitary pads, which students frequently cited as a major motivator to attend remedial classes. However, WUSC eventually replaced distributions of sanitary pads with cash transfers, because cash transfers could increase food security, and because evidence has demonstrated the positive effects of cash transfers on school attendance (Baird, Ferreira, Özler, & Woolcock, 2014).

Exhibit 11 summarises the findings related to gender norms.

Exhibit 11. Gender Norms: Findings and Strength of Evidence

Finding	Innovation Teams	Strength of Evidence From HEA
Sensitisations targeting parents and guardians helped to increase support for girls' education.	• WUSC	
Specialised support helped to reduce dropout amongst women and girls, and to reduce barriers to scaling.	• Kepler • WUSC	

II. Programme Business Models and Scale

This section analyses how the five innovation teams' business models—including organisational management, project design, and finances—affected their ability to scale up or scale out. The analysis identified the following overarching themes related to programme business models and scale:

- Conflict settings are complex and fragile, so it is imperative to have a solidly developed plan for implementation, and to develop a theory of change that draws on existing rigorous evidence from a wide range of settings, including development settings in low- and middle-income countries (and humanitarian settings, to the extent possible).
- Demand for many of the innovation teams' services outpaced their ability to design and establish formalised business systems, which should be in place prior to scaling. Formalised business systems help to smooth operations, while still leaving room for programmes to be flexible and adaptable (an inevitable and necessary part of the scaling process).
- Although our initial conceptual framework hypothesised that pilot programmes would scale up before scaling out to other contexts, most innovation teams followed a trajectory in which their programmes scaled *out* before scaling *up* due to strategic considerations, donor location priorities, and uncertainty about future funding.

This section discusses common findings related to five aspects of the innovation teams' business models: (a) organisational management, (b) financial resources, (c) implementation capacity, (d) project design and implementation, (e) non-governmental partnerships, and (f) exit strategy.

a. Organisational Management

Findings related to organisational management highlighted the need to formalise systems and documentation related to finances, personnel, design, and partnerships before scaling up or scaling out. This is consistent with the findings of Ramalingam and colleagues (2015), who argue that improving innovation management skills can help to facilitate scaling. Glewwe and Muralidharan (2015) found that interventions focused on improving management in education delivery were amongst the most effective for improving learning outcomes.

Photo Source: War Child Holland

Respondents across innovation teams in the pilot stage reported that formalised processes were the primary requirement for managing programmes on a larger scale. Each of the innovation teams referenced a tendency to focus on outward-facing end products, instead of prioritising the systems that would ensure high-quality implementation. As a result, almost all innovation teams encountered challenges related to management and implementation throughout programming, which they suggested could have been reduced through greater planning and coordination in the early stages of their projects.

Despite the need for formal systems, flexibility allowed for necessary adaptations to fit the context. Each of the innovation teams demonstrated a high level of flexibility to make their programme work as needed, depending on the context. LWB provided content identified through a needs assessment that responded to the interests and demands of people in the area. Kepler frequently adapted aspects of its programme as it learned more about the constraints that students faced.

One LWB respondent pointed to the need to maintain flexibility in the theory of change, especially given contextual differences:

> *The consequence of that [is being] locked inside a logical framework. … The flexible tool is meant to be reinvented once working with people [to generate human-centred design]. … There was frustration from [the] Ideas Box facilitator that she was restricted because she said she'd like to do [something different], but she can't.*

This frustration is similar to experiences documented in the *Doing Development Differently* manifesto, which argues for an adaptive and flexible approach to international development (Overseas Development Institute, 2014).

Exhibit 12 summarises the findings related to organisational management.

Exhibit 12. Organisational Management: Emerging Themes and Strength of Evidence

Finding	Innovation Teams	Strength of Evidence From HEA
Teams emphasised the need to formalise systems and documentation related to finances, personnel, design, and partnerships before scaling up or scaling out.	• Caritas • LWB • Kepler • WUSC	
Despite the need for formal systems, flexibility allowed for necessary adaptations to fit the context.	• Caritas • LWB • Kepler • WCH • WUSC	

b. Financial Resources

Each of the innovation teams had difficulty maintaining financing from limited and inconsistent grants that lasted for relatively short periods of time (many for 5 years or less). This problem has been identified in the existing literature on education programming in crisis settings (e.g., McClure & Gray 2015c; Results for Development & UNICEF, 2016; de Hoop et al., 2018). Respondents noted that having to continuously find funding through donor grant mechanisms—the primary funding source for all of the innovation teams—was neither efficient nor sustainable. Funds that are unreliable in the long term make planning to sustain the cost of scaling difficult. Teams also reported that decisions about scaling were made in response to donor priorities, including where to scale, and the extent to which it was possible to scale. As a result, teams were often unable to pursue longer term planning until funding was secured. Grant dependence created incentives to start multiple pilot projects in new contexts, rather than scaling up in the same context. For this reason, teams discussed transitioning to revenue-based funding.

It should be noted, however, that some teams started multiple pilot projects in new contexts for strategic reasons. WCH decided to start pilot projects in multiple contexts to develop a strong evidence base for the effectiveness of CWTL and to improve the codification or tangibility of the programme. Such strategic considerations may help in improving sustainability and moving the programme to scale.

> *"There are moments when we're not sure if we're going to continue doing what we're doing. That's important in terms of scaling—issues of financial resources."*

Some pilot programmes struggled to create scalable financial administration processes that remained relevant throughout the implementation period. Innovation teams reported that their processes for tracking finances were not consistently organised. For example, one respondent said that a lack of regulation on reporting amongst smaller donors meant that they were not required "to be super rigorous on expenses." Such gaps in financial systems can cause implementation delays as teams scale.

The innovation teams recognised the need to formalise financial reporting processes to ensure smooth financial operations (see Exhibit 13). Some teams in the scaling-out and scaling-up phases said that they

established more formalised systems and procedures for financial administration over time, after being externally incentivised to do so. The innovation teams suggested using existing information, resources, or professionals to coach teams on how to build functional financial systems or, alternatively, to incorporate financial tracking into regular M&E.

Volatile internal and external financial systems hampered programme implementation. In addition to challenges formalising their own financial systems, programmes' financial difficulties were compounded by the informal nature of many of the vendors with whom innovation teams worked. Many vendors had difficulty making electronic transactions, increasing the need for cash transactions, which are more likely to contain errors in documentation. Many vendors were also local, which meant that processes were sometimes unofficial, affecting financial tracking.

Lack of financial planning can also affect transactions with vendors and cause issues when employees do not receive payments on time. Some refugees who volunteered with Kepler said that payments could be delayed because refugee volunteers did not have bank accounts, which required Kepler to identify other formal, secure methods for transferring payments, such as money transfers. A staff member from Caritas reported that procurement for necessary classroom items was delayed because of internal procedures; in this particular instance, approval for payment had to go through as many as four levels. Three of the innovation teams also reported that their ability to use funds efficiently and appropriately was hampered by the need to approve transactions and disbursement at the central level. One team suggested that decentralising financial administration might facilitate payments to vendors and employees in the field.

Teams were considering ways to finance programme scaling beyond donors. All of the innovation teams indicated that continuing to rely exclusively on donor financing was not a sustainable approach to programme scaling because of the volatility of funds and the fact that donor demands varied by context. One respondent explained: "The landscape for philanthropy … is not sufficient enough to rely on it to make all of our scale goals." Finding additional sources of funding became a more salient challenge for teams as they scaled out, because short-term financing is a less viable option for maintaining a large programme. One team member spoke about the volatility of financing: "Funding is likely to continue for another year, [but the donor] is starting to phase out from [the country] because they don't see it as a humanitarian setting."

Given the challenges of securing government co-financing, especially in the initial stages of implementation, some teams were looking at the possibility of generating revenue by selling elements of their programme design or materials. Other teams, such as LWB, were using a shared-services model with other businesses operating in the same geographical areas. WCH was exploring subcontracting for a private-sector education initiative, and Kepler was considering ways to subsidise refugee education using non-refugee student fees. In most cases, teams are likely to continue looking to multiple sources to fund their work and solidify their approach.

Teams may be able to set up the same programme with lower costs when replicating their programmes in new contexts. A costing analysis indicated that Kepler may be able to reduce the costs of its programming when replicating its programme in a different context because it has already developed a curriculum in Kiziba. When scaling out to a different context, Kepler could achieve cost savings of close to

$66,000 (see Exhibit B-1 in Annex B). Similarly, a costing analysis by Jones (forthcoming) showed that WCH achieved significant economies of scale when moving to new contexts.

Exhibit 13 summarises the findings related to financial resources.

Exhibit 13. Financial Resources: Findings and Strength of Evidence

Finding	Innovation Teams	Strength of Evidence From HEA
Creating scalable funding and financial administration models from the outset was difficult.	• LWB • Kepler • WUSC	
Informal external financial systems sometimes delayed programme implementation.	• Caritas • Kepler • LWB • WUSC	
Teams were considering ways to finance programme scale beyond donors.	• Caritas • Kepler • LWB • WUSC • WCH	
Teams may be able to set up the same programme with lower costs when replicating their programmes in new contexts.	• Kepler • WCH	

c. Implementation Capacity (Personnel)

Findings from the meta-evaluation strengthen existing evidence that education programmes in protracted crisis settings struggle with inadequate capacity and challenging environments. Projects that lack human or material resources for effective implementation are likely to struggle during both the design and scaling-up phases. This section discusses challenges with implementation fidelity, including high staff turnover and a lack of clearly defined management roles and responsibilities.

Staff recruitment and retention initially affected training and implementation fidelity, but employing locals helped to increase retention. Three of the five innovation teams said that they initially experienced teacher turnover because of low pay and challenging living and working conditions. As a result of this high turnover, teams were not able to guarantee that all teachers had the same level of training and experience. This led to variation in the quality of programme implementation. One respondent from Caritas explained: "There is a regular turnover for the teachers—[it is] a huge investment of energy in explaining to the newcomers and informing the teachers about these methods." The respondent also indicated that additional training may not be worth the investment as teachers often stay for only 1 year.

Over time, most of the teams overcame staffing challenges, partly by adopting longer term teacher coaching models, which have been shown to be more effective than one-off teacher training (Cilliers & Taylor, 2017), and by relying on local staff to encourage ownership and retention. One respondent from WCH in Uganda noted: "Teaching assistants are qualified in South Sudan. They were brought into [the]

classroom to help with the number of students and to help with language/cultural issues and to support the students." Kepler regularly employed former students and worked with students' relatives as incentive-based volunteers who had a stake in the programme's success. Caritas, WUSC, and LWB also employed community members. Caritas's initial training was supplemented by regular, ongoing coaching.

Staff were in demand for work that went beyond their primary responsibilities. Three innovation teams reported that teachers, field staff, and management staff were responsible for work that extended beyond their primary roles. Respondents said teachers worked on non-classroom activities involving refugee legal issues, student–parent conflict, and community mobilisation, amongst others. A respondent from WUSC suggested that spending too much time mobilising community members to participate in the programme—which happened in both Kakuma and Dadaab—could harm teachers' performance in the classroom: "The teachers that are supposed to be working over the weekend also need time to prepare well. But if they've been spending a lot of time mobilising, it could be eating into their time."

Staff emphasised the need for clear management roles and responsibilities from the outset. Roles and responsibilities amongst management personnel were unclear and were not consistently documented in an organised and accessible way for all staff, usually from the beginning of implementation. This was the case for four of the five innovation teams, especially during the start-up phase. One respondent indicated that there were "start-up challenges. … How much time can we request from them? How to make sure they can do data uploads? How much do we need to support them? Staffing was all over the place and it was an issue." One team reported that a more consistent approach was needed to communicate programme policies to participants. In one case, participants said that they had noticed ongoing changes, which, at times, were sudden and negatively affected their experience with the programme. Two teams reported that their programmes lack formalised structures for management, especially during the pilot stage.

Staff identified a need for improved communication about implementation and management between global headquarters, country-level headquarters, and district and local levels. As with staffing, most of the teams were able to identify a need for stronger management structures and had developed solutions throughout implementation, including staffing meetings and efforts to clarify reporting structures. One respondent from WUSC explained:

> *When you are doing those proposals … we need to be very specific on who takes what. You see some other components are introduced that are not at the initial drafting of that particular proposal … but we can have those [planning] meetings; they make things clearer.*

However, multiple teams also indicated that communication between levels could improve, and that implementers could be more involved in planning and decision-making. One WCH team member at the country level stated:

> *[There is] a big gap between Kampala and the field. It is hard to make agreements at [the] Kampala level and see them happen in the field. We … need to look at how we support partners at the field level but make sure decision makers in Kampala are informed and agree. Want both sides to agree, but it is not always clear how they coordinate within the organisation itself.*

Exhibit 14 summarises the findings related to implementation capacity.

Exhibit 14. Implementation Capacity: Findings and Strength of Evidence

Finding	Innovation Teams	Strength of Evidence From HEA
Challenges to staff recruitment and retention initially affected training and implementation fidelity, but employing locals helped to increase retention.	• LWB • Caritas • WUSC	
Staff were in demand for work that went beyond their primary responsibilities.	• Kepler • LWB • WCH • Caritas • WUSC	
Staff emphasised the need for clear management roles and responsibilities from the outset.	• LWB • Kepler • Caritas • WUSC	
Staff identified the need for improved communication about implementation and management between global headquarters, country-level headquarters, and district and local levels.	• Kepler • LWB • WCH • Caritas	

d. Project Design and Implementation

Three innovation teams had difficulty defining core elements of their programme design. A respondent from LWB reported that there was no unified model of what the innovation should consist of in the field and suggested that lacking a core focus could hinder quality: "[We] struggle with having so many different objectives—it's hard to focus on quality [when we are] more focused on objectives." Another innovation team said that it was harder to scale to new contexts without a solid design. Being able to explain the core components of the programme, both initially and throughout implementation, could help teams to navigate other implementation challenges. This section discusses two areas in which teams consistently encountered challenges with implementation: obtaining permission to use space for their programmes, and fielding demands for services that went beyond the scope of their core programmes.

Teams made content and scaling decisions based on donor demands, which can affect planning for and the general ability to scale. Most of the teams reported that their decision-making about where to scale, and the extent to which they could scale, was based primarily on donor demand for services and funding priorities. For example, one team explained that it changed the focus of its programme to accommodate donor demand:

Photo Credit: WUSC

> *The decision … was mainly informed by donor requirement. They had these priorities: most marginalised, disabled, overaged learners. How do we move from where we are to apply for [a] call for proposals? We continued with [the original model], but with focus on overage and disabled to comply with donor requirements. [We] tweaked what we are doing to meet the needs of the funder.*

Teams indicated that this type of "timing and luck" model was not ideal for designing a programme at scale, given that implementation timelines and requirements vary by donor, which makes it difficult to create a single vision for a programme's core and wraparound services. One implementer also highlighted the challenge of short donor-funding cycles:

> *The project is only 3 months—not enough time to monitor and evaluate, and not easy to have these results in 3 months. [It is] not the result I was expecting to have if we had a full year of activities, [but we] couldn't extend the project because [there is] no funding.*

Teams identified some disconnect between theories of change and pathways that led to programme outcomes. Although it was important for teams to solidify their theories of change throughout the pilot stage of their programmes, most teams identified some gaps between the pathways in their theories of change and the potential to reach stated outcomes. As teams scale, it is helpful if the theories of change clearly reflect how each of the technical elements of a programme contributes to programme objectives (International Rescue Committee, 2019). For example, multiple teams mentioned the importance of continuous teacher training and mentoring—a component that was not explicit in all of the theories of change from the outset.

In at least one location, LWB had to create a logical framework that would fit within a partner's framework after implementation had already commenced. One respondent spoke about putting so much work into a new theory of change so close to implementation: "It's a symptom that the first conversations were limited. They were limited on our side, too." Limited early conversations about the theory of change may also have contributed to implementation challenges for the self-implemented Ideas Box in Jordan, where some English games were difficult to understand for children who only spoke Arabic. Facilitators had to work to find instructions in Arabic to use the games for sessions.

Procurement was delayed in locations where there was high demand for resources. Four of the innovation teams reported that they had underestimated the logistics required to set up programmes in a new location, and that this was further complicated by the cumbersome procedures for procuring goods that were in low supply. Without being able to procure the needed materials in time, implementation was delayed, and staff resources were diverted to logistical efforts. WCH said it understood procurement procedures only after scaling to a new location, where implementation was delayed as a result: "[We] also learned about procurement, which took a long time last year; now we know the suppliers and where to get quotations." A respondent from LWB reported that the programme started late in one location because of customs procedures: "When the box was in customs, it was delayed by administrative procedures. No specific reason—the release company did not communicate well with admin staff, the paper was not ready, something about the signature. It was about administrative procedures."

Exhibit 15 summarises the findings related to project design and implementation.

Exhibit 15. Project Design and Implementation: Findings and Strength of Evidence

Finding	Innovation Teams	Strength of Evidence From HEA
Teams had trouble defining the core components and wraparound services for their innovations well into the period of programme implementation.	• LWB • Caritas	
Teams made content and scaling decisions based on donor demands, which can affect planning for and the general ability to scale.	• LWB • Caritas • WUSC • WCH	
Teams identified some disconnect between theories of change and pathways that led to programme outcomes.	• Caritas • Kepler • WUSC • LWB	
Procurement was delayed in locations where there was high demand for resources.	• Caritas • Kepler • LWB • WCH	

e. Non-Governmental Partnerships

Findings indicated that strong partnerships were essential to running even minor aspects of programming. Innovation teams relied on partnerships to overcome challenges related to security, physical resources, access, and many other aspects of implementation, and teams recognised the need for a more strategic approach to partnerships from the outset in order to facilitate these operations. One team explained: "The core to strong partnership is having values and mission alignment from the beginning and exploring enough with the partner to know that you do." This section discusses two challenges related to partnerships that were consistent across multiple innovation teams: the difficulty of working within a large network of service providers, and the disadvantages of having informal partner agreements. We then discuss how networks of strong, on-the-ground partners can facilitate implementation.

Pilot programmes encountered difficulties as they tried to fit within an intricate network of actors. Despite the existence of complicated networks of NGO actors in emergency settings, four of the innovation teams said that they needed to move away from the current "ad hoc" approach to establishing partnerships. A respondent from the WUSC team illustrated the complex network of actors in emergency settings by describing a sample of the partners they work with to implement programming:

> *We have Windle, obviously for secondary education; they also implement some high education initiatives and such projects as the DFID-funded girls' education project [and] the BPRM [Bureau of Population, Refugees, and Migration] remedial training. We also have LWF [Lutheran World Federation]—they implement primary education and pre-primary education,*

> *and special needs education. They are also involved in teacher training. Then we have Don Bosco, which does vocational training. NRC [Norwegian Refugee Council] is also involved in youth education programmes and trainings.*

A team member from LWB suggested using larger partnership frameworks rather than individual partnerships as they scale the programme: "[We] need now to sign larger partnerships, not one by one; we should focus on creating larger partnerships in one or several camps—larger partnerships with NGOs or agencies like UNHCR." Another respondent pointed to the need to think about expansion in terms of the partnerships needed to cover programme implementation at scale: "A lot of [the programme is] going on, but it's more divided among partners than we had originally thought. … We thought more organisations [running the programme] would be [larger], but the majority of organisations are smaller." Although the team said they did not reach their original target number of geographic locations, they had "[a] better understanding of education in emergencies" with regard to how partnerships affect scaling capacity. One team also recommended having "a rubric for identifying partners, a midpoint check-in with partners that indicates how well do you think you're doing, how are these systems working for you, where things need to change, sharing with partners student needs as first."

Informal partner agreements left elements of programme implementation volatile. Multiple innovation teams indicated the importance of establishing parameters for a partnership, even during the pilot stage. However, this practice was lacking amongst four of the innovation teams. Formalising agreements is especially challenging in humanitarian contexts, given the presence of competing programmes, difficulties accessing certain areas, and the additional approvals needed to operate. One respondent from LWB emphasised the need to ensure that both parties were clear on who was responsible for each aspect of the programme, as well as any reporting requirements. Other suggestions for formalising partnerships included systematically selecting partners based on requirements in a given context, establishing larger partnership frameworks, and maintaining an in-country presence.

Networks were a primary means of establishing and maintaining strong, on-the-ground partnerships. Four innovation teams in the scaling-out phase reported that networks of trustworthy and experienced partners helped to establish additional programme sites and eventually took over implementation. In some contexts, WCH was shifting away from active involvement in on-the-ground implementation and was instead working through local, experienced partners. Caritas has been able to scale out to new locations as emergencies arise because of networks it established through EoL and other Caritas programmes. LWB reported that a strong existing partner (the International Rescue Committee) was able to take over programme management in some settings.

Teams identified a need to better coordinate across sectors to streamline programming, reduce overlap, and make better use of sector-specific expertise. During the final round of data collection, three innovation teams highlighted the importance of cross-sector collaboration to streamline programming and lower costs. On topics peripheral to education—such as psychosocial support, cash transfers, and other areas of social protection—programmes had difficulty finding experts to ensure the highest quality implementation of those programme components. One respondent commented: "Instead of competing for the funding, why don't we apply together … and try to harmonise implementation?" Another respondent from WUSC suggested that it was important to "make sure managers link what is being done [outside of school] to the needs and gaps of regular school." Most of the teams had already begun cross-

sector collaboration and efforts to reduce redundancy within education programming. One respondent suggested that cross-sector partnerships could ensure appropriate use of available expertise and streamline funding and implementation.

Exhibit 16 summarises the findings related to partnerships.

Exhibit 16. Partnerships: Findings and Strength of Evidence

Finding	Innovation Teams	Strength of Evidence From HEA
Pilot programmes encountered difficulty when trying to work within an intricate network of actors in emergency settings.	• LWB • Kepler • Caritas • WUSC	
Teams used networks to establish and maintain strong, on-the-ground partnerships.	• Caritas • Kepler • LWB • WCH • WUSC	
Informal partner agreements left elements of programme implementation volatile.	• Kepler • LWB • Caritas • WUSC	
Teams identified a need to better coordinate across sectors to streamline programming, reduce overlap, and make better use of sector-specific expertise.	• WUSC • Caritas • WCH • LWB	

f. Exit Strategy

> *"One of the communications concerns the sustainability; it's about whether we have the capacity to sustain and continue using the Ideas Box in a productive way. Having the Ideas Box without the smart, talented, creative people—it will turn like any laptop. We lack human resources, talented, creative human resources."*
>
> *— LWB Respondent*

Programme sustainability depends on many of the factors discussed in this chapter—including funding sources, relationships, and alignment with partners and governments—from the pilot stage onward (i3, 2017; Melaville, Jacobson, & Blank, 2011). Research also suggests that organisations should ideally establish a vision for scaling during the pilot phase (Hartmann & Linn, 2007; Melaville, Jacobson, & Blank, 2011; Panirsilvam, 2017). However, in a review of 45 different, successfully scaled programmes, Larson, Dearing, and Backer (2017) found that most organisations made no mention of scaling plans during the pilot phase.

Four of the teams considered the need to create a model to transition programme operations to a local or government entity, although the mechanism for doing so varied by location and context. A respondent

from LWB illustrated the difficulty of sustaining a programme without a solid plan and reliable partners, which ideally should be in place from the outset: "In Greece, there's a project that's ended, and the Ideas Boxes are locked in containers and not being used." Innovation teams discussed approaches to financing primarily or exclusively through local partners, although not until well past the pilot stage.

Teams did not incorporate plans to sustain programming from the outset, including longer term funding and transition strategies. Most of the five innovation teams had an established vision for longer term sustainability or contemplated initiating such discussions at the beginning of their programmes. As a result, teams encountered challenges in maintaining funding for the pilots and subsequent funding for implementation on a larger scale. In various locations, respondents spoke of uncertainty about continued funding or the ability to transition the project to a local partner. A respondent from one organisation explained that "most people are not thinking past 2022" (the end of the funding period). All of the teams engaged in discussions with governments from the outset. However, conversations about how governments could help sustain programming were limited and did not concretely propose ways in which governments could contribute funding.

Financing or working through partners is a promising approach to sustainability and scaling. Multiple teams alluded to the idea that sustainability depends on partners eventually being able to finance and run a programme independently. Teams aimed to incorporate a strategy in their visions for scale for finding new partners and developing local systems in which partners could finance parts of programming. For example, Kepler's degree-granting partner, SNHU, is starting to scale out, using the online platform with different implementing partners in other settings. One respondent from SNHU said that while the core elements of the programme remained the same across partners in different locations, "lots of logistical pieces," including management and logistics, became important elements during scaling.

Local implementation and procurement can also bring down cost by allocating responsibilities to local partners. One respondent explained the advantage of partnering on implementation: "It's too expensive … particularly because we still have conversations that are unisectoral. So, we are having conversations with partners looking to reach a small target group."

Exhibit 17 summarises the findings related to exit strategies.

Exhibit 17. Exit Strategy: Findings and Strength of Evidence

Finding	Innovation Teams	Strength of Evidence From HEA
Teams did not incorporate plans to sustain programming from the pilot stage, including longer term funding and transition strategies.	• Kepler • LWB • WUSC • Caritas	
Financing or working through partners is a promising approach to sustainability and scale.	• LWB • WCH • WUSC	

III. Community Support, MoE Engagement, and Scaling

Recognising that a lack of community support can act as a barrier to scaling interventions in protracted crisis settings (Agha, 2016; Li, 2012), this section examines how HEA programmes fostered community support, ownership, and demand for their interventions. Existing evidence demonstrates the importance of close government collaboration for effective education programming at scale (de Hoop et al., 2018). The latter portion of this section explores lessons learned about engaging with MoEs and securing political buy-in for HEA programmes.

Research Questions

- How have HEA programmes fostered community support for their interventions?
- What worked to increase demand for HEA programmes and ensure buy-in at all levels?
- What lessons have innovation teams learned about engaging MoEs successfully?

a. Community Support and Ownership

Innovation teams considered transparency and regular communication with stakeholders essential to build trust and establish support for programmes at the community level. All five innovation teams emphasised the importance of building trust and support through clear and regular communication. This aligns with the literature, which highlights the importance of engaging end users and gatekeepers to foster community-level support for innovations to facilitate scaling (Obrecht & Warner, 2016), as well as the importance of two-way communication between service providers and crisis-affected people (Brown & Donini, 2014; Chapelier & Shah, 2013).

WCH informants discussed the value of continued communication throughout programme implementation, particularly when faced with delays. In Sudan, for example, CWTL staff visited villages to explain implementation delays and provide updates. Caritas informants emphasised the importance of building trust (particularly with parents) by ensuring that communities understood exactly what the EoL programme was and what its objectives were. Respondents from Kepler and LWB underscored the need for transparency and frequent communication to maintain positive perceptions and community support for their programmes. Key informants from WUSC also emphasised the importance of regular community engagement, not only to support the remedial programme but also to support schools themselves and elevate the importance of girls' education. WUSC's approach to community engagement aligns with Rose and Greeley's (2006) recommendation to prioritise the institutionalisation of school–community relationships for education programmes in fragile states.

> *"Then we also work closely with the community because we realised that you can't do it alone. You know to them … they listen more to their community members, particularly the leaders, their parents, the head teachers of those schools. And some schools even [if] the head teacher tells them don't go to the remedial, they don't go. So, you must cultivate a very good working relationship, cordial relationship with the head teacher."*
>
> *— WUSC informant*

Targeted communication through influential actors (gatekeepers) built community support for HEA programmes and generated demand. All five innovation teams used regular communication and targeted outreach to encourage demand for their innovation, relying in many cases on influential community members (or gatekeepers) who played a

significant role in generating demand from the ultimate end users: the children learning through these interventions. Innovation teams engaged gatekeepers to perform targeted communication and outreach in slightly different ways. WCH brought together head teachers, teachers, parents, and community members to explain the CWTL programme before it began, while WUSC engaged community mobilisers to target certain blocks (and block leaders) in Dadaab and Kakuma for remedial recruitment and attendance. LWB worked with community-based organisations and held events to reach out to potential Ideas Box users, while Caritas worked with community liaisons or community facilitators to target parents and community members in Satu Mare to increase demand for EoL. Kepler provided targeted outreach to students through the secondary schools in Kiziba camp, as well as through meetings with their parents. In each of these cases, using respected and influential actors to transmit information about education interventions garnered support and encouraged participation.

Innovation teams built ownership through co-development of content, involving key stakeholders in activities. Obrecht and Warner (2016) highlight the importance of ownership to achieve the ultimate goal of innovation "diffusion." Two primary approaches that encourage programme ownership emerged across the five innovation teams: co-developing programme content, and involving key stakeholders (such as parents or former participants) in programme activities. The importance of content co-development prior to programme roll-out came across clearly in the LWB data, while the involvement of key stakeholders in actual programme activities was more apparent for Caritas, Kepler and WUSC. For WCH, both co-development and key stakeholder involvement were important.

The Brookings Institution (2008) highlights the importance of consulting people affected by conflict in programme design for both "normative and instrumental reasons." In developing content for their programmes, WCH and LWB both incorporated elements of user-centred design—an approach particularly relevant when developing an innovation for conflict-affected people (Obrecht & Warner, 2016). WCH believes that involving key stakeholders from the early design stages, and in content co-creation, is critical to ensuring relevance and ownership of the programme: "Co-creation ensures ownership from the beginning and ensures content remains relevant."

Photo Source: Caritas Switzerland

LWB's needs assessments followed similar principles, encouraging future users of the Ideas Box to suggest content and build relationships with staff before activities officially started. Perhaps partially due to this early engagement with communities, LWB has been particularly successful in promoting ownership of Ideas Boxes in some settings. For example, a key informant from LWB described a situation where people protected the Ideas Box from vandalism: "Some buildings were set on fire within camp, [but] some of the users within the camp started standing around Ideas Box, protecting it and saying, 'This is useful for the community and we shouldn't destroy it.'" In Uganda, teachers and programme staff reported that students and surrounding communities felt that the CWTL tablets and solar panels were theirs (not the property of the implementing NGO) and attributed the absence of damage or theft

to this sense of ownership, which they helped to build through early meetings with parents and nearby community members.

While WCH and LWB involved end users in the content development process, the interventions themselves (tablet-based educational gaming and a mobile library) were predetermined—an aspect of the programmes that stands at odds with the core principle of human- or user-centred design. A truly human-centred approach would involve entering a context without a predetermined intervention or solution and engaging end users to collectively develop a solution (IDEO, 2015). While the content within all five HEA interventions was tailored to the needs of the end users, those users were not consistently involved in the initial conceptualisation of the interventions.

Implementers from Kepler encouraged programme ownership by employing parents and former students. Kepler also adjusted curriculum content and programme offerings between baseline and endline in response to participant feedback. In Romania, Caritas social centre staff actively engaged parents (often through community liaisons) and promoted ownership of the EoL programme by allowing the centre to serve as a de facto community centre for the Roma community in Satu Mare. In Dadaab and Kakuma, WUSC relied on community mobilisers to support both EERCK and KEEP programme activities. WUSC respondents attributed the involvement of EERCK refugee staff to increased community ownership: "If [the] community owns the project, it has some life, because they feel it is their programme … they can run with it … considering EERCK is a project that is [mostly] managed by refugee staff."

Adapting programmes to meet local needs increased their relevance and enhanced community support. Existing implementation evidence points to the need for flexibility and adaptation when scaling education programmes in humanitarian settings (Easton-Calabria & Omata, 2016; UNICEF, 2013). All five innovation teams made programmatic adaptations to better meet local needs, which not only made the interventions more relevant but also simultaneously increased community support for the programmes. Not all adaptations were based on empirical evidence, however, nor were they specifically made to facilitate the achievement of outcomes listed in the programmes' theories of change. Kepler made one of the few adaptations that was based on empirical evidence, in response to the need to demonstrate to the Kiziba community that its programme resulted in employment for its students. One Kepler team member said that recognising the need for more concrete student outcomes led to a push for improvements in the employment component, such as having internships occur earlier in the programme. Changing the timing of the internships was directly tied to Kepler's theory of change, which listed employability as an intended outcome.

Caritas realised that it needed to make programmatic adjustments to maximise parental support for the EoL programme in Romania. Social centre staff in Satu Mare adapted their approach by emphasising elements of the centre that parents believed were most useful, including homework help and material supports such as food, clean clothes, and showers. Caritas incorporated homework support because, according to several interviewees, ensuring that students completed their government school homework at the centre helped to foster parental support for the programme. In both Romania and Bangladesh, Caritas teachers were encouraged to make their own contextually appropriate learning materials and curriculum, based on the primary needs of students in those contexts. In Kakuma and Dadaab, WUSC liaised with communities to adjust the timing of remedial classes in accordance with local religious obligations. WCH in Uganda and LWB in Jordan made slight programmatic adjustments to ensure that

content was appropriate for the age and skill levels of participants. In Uganda, this meant changing the primary school grade in which the CWTL programme was implemented; in Jordan, it meant revising the content of the DRC Ideas Box so that it was suitable for a younger audience.

Exhibit 18 summarises the findings related to community support and ownership.

Exhibit 18. Community Support and Ownership: Findings and Strength of Evidence

Finding	Innovation Teams	Strength of Evidence From HEA
Transparency and regular communication with stakeholders were essential to building trust and establishing community support for programmes.	• Caritas • Kepler • LWB • WCH • WUSC	
Targeted communication through influential actors built community support for programmes and generated demand.	• Caritas • Kepler • LWB • WCH • WUSC	
Co-development of content and/or involving key stakeholders in activities built community support and demand for programmes.	• Caritas • Kepler • LWB • WCH • WUSC	
Adapting programmes to meet local needs increased their relevance and enhanced community support.	• Caritas • Kepler • LWB • WCH • WUSC	

b. MoE Engagement

Innovation teams engaged with MoEs early in the pilot phase, but this did not always include plans to scale up or institutionalise interventions. Existing evidence shows that close collaboration with local governments is required to effectively implement education programmes at scale (de Hoop et al., 2018). All five innovation teams noted the importance of securing political buy-in and government endorsement (particularly from MoEs) for their programmes early on, with informants from CWTL repeatedly mentioning the need to secure MoE approval and buy-in early and throughout the process of expanding to a new country. However, programmes' early engagement with local governments did not necessarily include a discussion of longer term plans to bring interventions to scale. For example, Kepler and Caritas aimed to refine their programmes for each context through pilots before engaging in conversations about scaling, institutionalisation, and a continuation of financing for their programmes.

Demonstrating adaptability and customisation facilitated MoE buy-in. During initial discussions with MoEs, WCH said that it was important to demonstrate how the CWTL programme would be adapted in close collaboration with the MoE to ensure appropriateness in terms of look, feel, and content. WCH believes CWTL's success is due in part to tailoring content to each context and meeting the needs and priorities of the MoE. WCH underscored the importance of fully understanding the education situation in the country, as well as the motivation and strategy of the MoE, before beginning conversations. In Lebanon, for example, WCH recognised the MoE's emphasis on getting all children (including refugee children) into school and was able to demonstrate how the CWTL programme would help them do that. WCH staff also highlighted the importance of differentiating the CWTL programme from others and emphasising that it was more than "just a game" or "just a tablet."

Having evidence of programme effectiveness and a strong reputation helped to secure government buy-in. The literature suggests that generating evidence of effectiveness can help programmes scale up (de Hoop et al., 2018). Three of the five innovation teams (LWB, WCH, and WUSC) also emphasised the importance of collecting evidence of programme effectiveness and developing a strong reputation when approaching potential government partners. CWTL defined evidence of effectiveness as rigorous research on well-being outcomes, user experience, and learning outcomes in the given subject (reading or mathematics) in the local language. Other teams noted the value of a strong reputation. One LWB respondent explained that the process of garnering support from local authorities took time as LWB worked to prove the utility of the Ideas Box: "Now we're known by federal organisations, federal NGOs. They know what services we provide, what [the] quality of our work is. It can take time." In thinking about scaling EERCK, a WUSC respondent referenced "getting the word out … showing we have evidence" as a prerequisite for political buy-in.

Government relationships required maintenance and could evolve over time due to changing policies and priorities. Close engagement with government contributes to effective programme implementation at scale (de Hoop et al., 2018). Four innovation teams (Caritas, Kepler, WCH, and WUSC) reported that they made concerted efforts to maintain relationships with government entities, manage expectations, and stay abreast of current priorities. Caritas staff in Romania explained that EoL staff worked hard to maintain strong partnerships with public school staff and local MoE staff by interacting regularly, discussing student challenges in public schools, and talking to school principals about student performance. WCH emphasised the importance of actively managing relationships with ministries, as well as managing their expectations of the CWTL programme. Having worked with multiple ministries on the same intervention, WCH observed that each ministry was different and that each relationship required a different approach.

> *With comprehensive refugee framework and [the] need for integration and inclusion in the context where we're working, I'd say there's an increase in terms of engagement with ministry. Because there's an appreciation that Kenyans are actually benefitting from education interventions in the camps.*

For WUSC, development of the comprehensive refugee framework in Kenya shifted its relationship with the MoE:

While the Kenyan MoE had previously adopted a more hands-off approach to refugee education (largely driven by UNHCR and international NGOs), WUSC informants maintained that the comprehensive refugee

framework shifted the MoE's thinking, and that its interest in and engagement with WUSC's remedial programming increased as a result.

Exhibit 19 summarises the findings related to political buy-in and integration into local education systems.

Exhibit 19. MoE Engagement: Findings and Strength of Evidence

Finding	Innovation Teams	Strength of Evidence From HEA
Early engagement with MoEs (during the pilot phase) was essential when expanding to a new country/context.	• Caritas • Kepler • LWB • WCH • WUSC	
Early engagement with MoEs did not always include plans to scale up or institutionalise interventions.	• Caritas • Kepler	
Demonstrating programme adaptability and customisation to meet local needs helped to secure MoE buy-in.	• WCH	
Providing evidence of programme effectiveness helped to secure MoE buy-in.	• LWB • WCH • WUSC	
Government relationships required maintenance and could evolve over time due to changing policies, priorities, and personnel.	• Caritas • Kepler • WCH • WUSC	
When incorporating an innovative education intervention into an established curriculum, the timetable required to deliver the intervention properly (while still adhering to the required content) sometimes proved challenging.	• WCH	

Synthesis on the Impact of Innovations in Education

This synthesis on the impact of innovations in education brings together lessons learned from impact evaluations of three innovation teams and the global literature on the impact of education innovations.

Impacts on Educational Outcomes

The quantitative synthesis found limited evidence that education innovations were having positive effects on learning outcomes amongst in-school children, possibly because of several implementation and contextual challenges.[6] Each of these challenges can be linked to the World Bank's conceptual framework, introduced in Section 3. Implementation and contextual challenges may have contributed to a lack of

[6] Annex I summarises the methodology for the individual impact evaluations.

positive effects for both remedial education programmes in Kakuma and Dadaab[7] and curriculum-based gaming approaches using tablets in Jordan.[8] Importantly, however, it is also possible that the programmes may have more positive effects in the longer term. Our evaluations only focused on relatively short time frames, during which it may have been challenging to achieve positive impacts on learning outcomes.

Challenges with learners and teachers may have limited the effects of remedial education on learning outcomes. Existing evidence suggests that remedial education programmes are effective in improving educational outcomes in some contexts, but they have not consistently shown positive effects on learning outcomes in low- and middle-income countries. A meta-analysis focused on three different remedial education programmes in Chile, India, and Mexico (Banerjee, Cole, Duflo, & Linden, 2007; Cabezas, Cuesta, & Gallego, 2011; Gutiérrez & Rodrigo, 2014) showed that remedial education, on average, did not yield statistically significant effects, after removing the study by Lakshminarayana and colleagues (2013) from the analysis (Snilstveit et al., 2016).[9]

Although remedial education is a promising intervention (Snilstveit et al., 2016), impact evaluations of WUSC's remedial education programmes in Kakuma and Dadaab refugee camps did not find statistically significant effects on Kenya Certificate of Primary Education (KCPE) exam scores. Participants in WUSC's EERCK programme did not score statistically higher on English, Kiswahili, mathematics, social science, or science outcomes than the control or comparison group in either Kakuma or Dadaab. We also did not find statistically significant effects on learning outcomes measured by the Uwezo reading assessment—a short assessment developed to evaluate the basic reading skills of children in Kenya, Tanzania, and Uganda (Uwezo, 2015).

Several implementation challenges may have contributed to the EERCK programme's lack of positive impacts on learning outcomes. Perhaps most importantly, students in Kakuma attended remedial education irregularly. Exhibit 20 shows the distribution of remedial attendance for girls participating in the programme in Kakuma. Of the eligible students, only 34% attended at least 50% of the remedial classes, and just 13% attended at least 75% of the classes. On average, girls attended 56% of the offered class hours in remedial education in Kakuma. One remedial teacher commented:

> *There was a day like on Saturday I was teaching 10 girls, and the next day I have introduced one chapter like how water is polluted. Then the next day, the 10 girls are absent; then other faces come. So they were like … floating; they could not understand anything.*

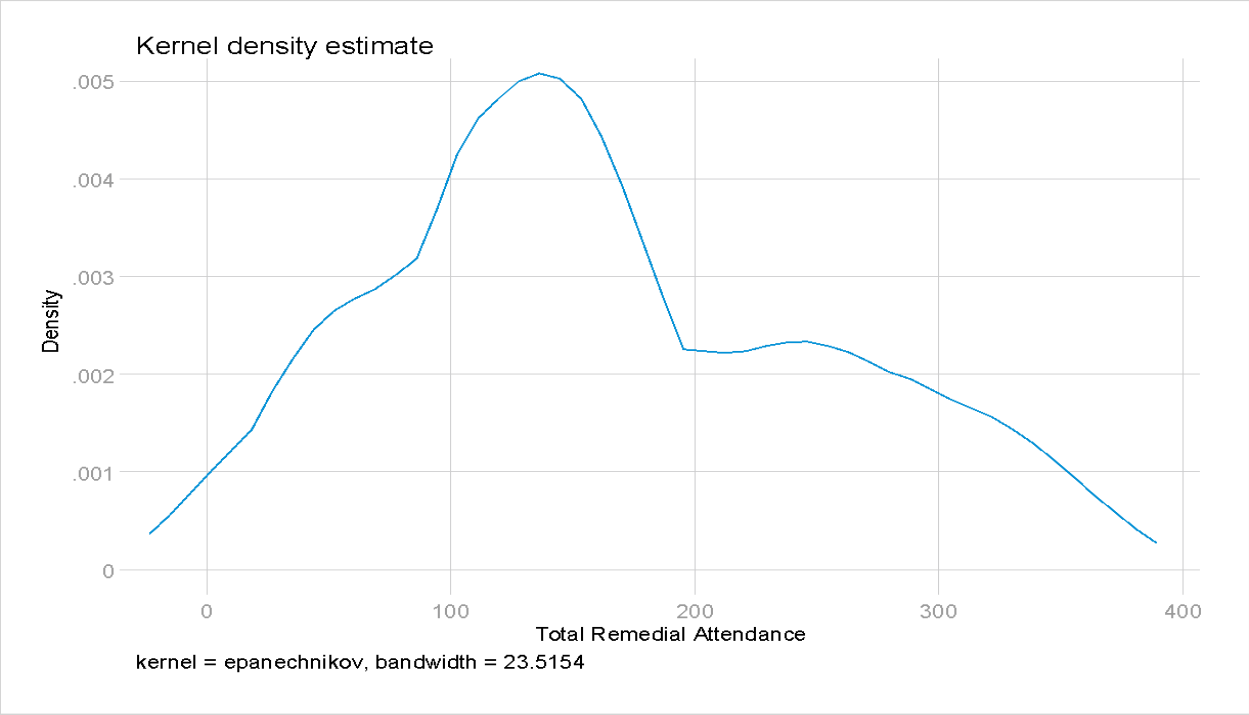
Remedial attendance was slightly higher in Dadaab, where administrative data indicated that girls attended an average of 61% of offered class hours.

The qualitative results showed that WUSC's remedial education programme also faced challenges associated with high teacher turnover, which may have limited programme effectiveness. As a result of this turnover, WUSC was not able to guarantee that all teachers had the same level of training and experience prior to implementation. In addition, only 26% of the students in Kakuma and Dadaab lived in

[7] Annex D presents impact estimates for WUSC's remedial education programme (EERCK) in Kakuma and Dadaab.
[8] Annex C presents impact estimates for WCH's CWTL programme in Jordan and Sudan.
[9] In line with Brody and colleagues (2015, 2017) and Chinen and associates (2017), we report the effectiveness of remedial education programmes after removing this study because it is a clear outlier that shows larger effects but has a high risk of selection bias.

food-secure households. This may have created educational challenges, given that food security is a key condition for improving learning outcomes (Snilstveit et al., 2016).

Exhibit 20. Distribution of Remedial Attendance



Challenges with learners and teachers limited the effects of technology-in-education programmes on learning outcomes. Bulman and Fairlie (2016, p. 2) report that studies that have looked at the impact of technology-in-education programmes on learning outcomes show "mixed evidence with a pattern of null results." However, recent evidence shows more promising results for programmes that include a strong focus on teaching at the right level. For example, Muralidharan, Singh, and Ganimian (2019) showed that a technology-based afterschool instruction programme with a strong emphasis on learning at the right level produced large and statistically significant effects on reading outcomes. A meta-analysis on the impact of technology in education also suggested that technology-in-education programmes can be effective if accompanied by parent or teacher training (Barrera-Osorio & Linden, 2009; McEwan, 2015). The analysis showed that technology-in-education programmes produced effects of 0.15 standardised mean differences on learning outcomes, on average, but that these effects became non-significant when the programme was not accompanied by parent or student training. Technology-in-education programmes can have negative effects on learning outcomes if they are not accompanied by additional programme components that aim to improve pedagogical practices (Stone et al., 2018).

Our analysis found that WCH's CWTL programme in Jordan produced equal learning gains in Arabic reading and mathematics amongst children receiving CWTL 40% of the time (and the standard government curriculum 60% of the time) and children receiving the standard government curriculum. We also did not find that the programme had differential effects on the Arabic reading assessment sub-components or the mathematics assessment sub-components amongst children receiving CWTL 40% of the time and children

receiving the standard government curriculum. As with the WUSC remedial education programmes, this lack of positive effects relative to the comparison group can potentially be explained by four operational challenges identified through the qualitative analysis: (1) limitations of the teacher training, (2) equipment malfunctions, (3) time management difficulties in the classroom, and (4) boredom associated with repetition within the game.

Addressing implementation and contextual challenges may result in larger effects on learning outcomes for education innovations. The quantitative synthesis indicated that addressing some of the implementation and contextual challenges may increase the effects of remedial education programmes on learning outcomes. For example, exploratory evidence suggests that WUSC's EERCK programme could have positive effects on learning outcomes for girls in food-secure households who attend remedial education classes more regularly in Kakuma. Impact estimates for the sub-sample of food-secure households showed that girls who lived in those households and attended at least 50 hours of remedial education in the year before the survey was conducted scored 0.20 standardised mean differences higher on the KCPE exam than girls in food-secure households in the control group.[10] This effect increased further for girls who attended more than 50 hours of remedial education per year.

These results suggest that the effects of WUSC's remedial education programme may be improved by incorporating a supplemental food security component or a social protection programme that has been shown to increase food security, such as cash transfers (Gilligan et al., 2013). Evidence shows that cash transfers can improve educational outcomes in refugee settings (de Hoop et al., 2019), and a systematic review has demonstrated that school feeding programmes have positive effects on learning outcomes. A meta-analysis showed that school feeding programmes increased composite standardised assessment test scores by 0.14 standardised mean differences (Snilstveit et al., 2016).

However, it is important to exercise caution when interpreting these results. First, we did not find statistically significant impacts for girls from food-secure households who attended more remedial education classes when we used the reading and mathematics scores from the Uwezo test as an outcome measure. Second, the exploratory results on the heterogeneous effects for girls in food-secure households who attended remedial education classes more regularly were based on a sample of just 170 students.

Remedial education did not have positive effects on school attendance. Both the global evidence and the impact evaluations of WUSC's remedial education programmes in Kakuma and Dadaab did not show statistically significant effects on school attendance. Access to remedial education did not affect primary school attendance in either Kakuma or Dadaab.

There was mixed evidence regarding the effects of technology-in-education programmes on school attendance. Our quantitative evidence did not show that WCH's CWTL programme had positive and statistically significant effects on school attendance in Jordan, for either boys or girls. However, it is important to exercise caution when interpreting these results because the introduction of the CWTL programme coincided with the introduction of more rigorous measurement of school attendance in treatment schools. Qualitative evidence suggests that school attendance may have increased as a result

[10] The impact estimates were statistically significant but were not robust to adjusting for multiple comparisons.

of the programme because the tablets improved perceptions of the quality of education in Jordan, Lebanon, and Uganda.

Multifaceted technology-in-education programmes had meaningful and positive effects on learning outcomes amongst out-of-school children and may be cost-effective. The impact evaluation in Sudan addressed a significant evidence gap by demonstrating that WCH's CWTL programme had positive effects on out-of-school children in Sudan.[11] To the best of our knowledge, the impact evaluation of the CWTL programme in Sudan is the first impact evaluation of a digital, game-based learning programme that focuses on learning outcomes for out-of-school children. The results showed that WCH's CWTL programme had considerable and statistically significant impacts on students' Arabic and mathematics outcomes, with impact estimates suggesting a positive impact of 0.78 standardised mean differences in Arabic reading outcomes and 0.63 standardised mean differences in mathematics outcomes. Although we are not aware of studies that show learning gains over the course of a school year in Sudan, the learning gains are equivalent to 3.7 years of education for Arabic reading and 3.0 years of education for mathematics (Evans & Yuan, 2019). These learning gains suggest that WCH's CWTL programme could be a cost-effective way of achieving improvements in learning outcomes for out-of-school children in Sudan. A simulation analysis showed that costs per programme participant can decrease significantly because of economies of scale (Jones, forthcoming).

Research also provided evidence of improvements in learning outcomes for out-of-school children in Lebanon and Jordan. In a practice-based evaluation, we found promising positive trends in mathematics outcomes for out-of-school children who participated in the CWTL programme in Lebanon, which key informants attributed to the CWTL programme. We also found promising positive trends in Arabic reading and mathematics outcomes for out-of-school children who participated in LWB's self-managed Ideas Box in Jordan (particularly as average scores in Arabic and reading outcomes went down in comparison community centres). Although neither of these studies assessed programme effects relative to a comparison group, the results were consistent with the positive impacts of WCH's CWTL programme on reading and mathematics outcomes for out-of-school students in Sudan.

Teacher coaching produced larger effects on learning outcomes than teacher training and may be cost-effective in improving learning outcomes. Although most innovation teams managed to adopt longer term teacher coaching models, the qualitative synthesis showed that several innovation teams initially struggled to find sufficient time and resources to provide teacher professional development. For example, WUSC found it challenging to schedule time for teacher professional development due to high teacher turnover, and several respondents suggested that Kepler should continue to strengthen its teacher professional development. However, Caritas managed to create a practical teacher training programme that enabled teaching of the activity-based curriculum, even to illiterate women; and remedial teachers in Dadaab and Kakuma largely had positive perceptions of WUSC's teacher training. Most of the teams also overcame staffing challenges by adopting longer term teacher coaching models.

Meta-analyses found mixed evidence on the effects of teacher professional development on learning outcomes in low- and middle-income countries. Popova, Evans, and Arancibia (2016) note that there is limited rigorous evidence on the effects of in-service teacher training, despite nearly two thirds of World

[11] WCH only recently finalised the analysis for the impact evaluation in Sudan. AIR is in the process of replicating the results.

Bank-funded education programmes including a professional development component. Findings from a systematic review of what works to improve early-grade literacy in Latin America and the Caribbean suggested that teacher coaching was more effective than teacher training (American Institutes for Research, 2016b). Similarly, Cilliers and Taylor (2017) showed that monthly visits from specialised training coaches resulted in large and statistically significant effects on reading outcomes (0.25 standard deviations) in South Africa, while two 2-day training sessions (provided over the course of a year) resulted in statistically insignificant impacts of only 0.11 standard deviations. These findings suggest that one-off investments in teacher training may only change knowledge, while teacher coaching focuses on developing a skill through ongoing practice (Cilliers & Taylor, 2017). These results are consistent with findings from Liberia, Kenya, and Uganda, where a programme that combined training, lesson plans, and coaching had large and statistically significant effects on early-grade reading outcomes (Piper, Zuilkowski, & Mugenda, 2014; Piper & Korda, 2011; Lucas, McEwan, Ngware, & Oketch, 2014).

It may be challenging for innovation teams to invest in teacher coaching, given the budgetary challenges and high teacher turnover. However, Cilliers and Taylor (2017) showed that teacher coaching is more cost-effective than teacher training in all scenarios: teacher coaching resulted in a 0.41 standard deviation increase per $100 spent per pupil, while one-off teacher training resulted in only a 0.23 standard deviation increase per $100 spent per pupil. Although we cannot extrapolate these results to complex emergencies, it is plausible that teacher coaching could result in larger impact estimates in protracted humanitarian crisis settings, as long as teacher turnover is limited.

The education innovations found mixed evidence on the impact of the programmes on learning outcomes when integrated into national education systems, possibly because teachers faced time constraints when trying to cover new material in addition to the prescribed national curriculum. The evidence synthesis on the impact of education innovations on learning outcomes reinforced that it is challenging to produce positive impacts when integrating education programmes into national education systems. For example, evidence from Kenya showed that assigning an additional contract teacher to tutor struggling students was effective in improving learning outcomes when implemented by an NGO during a pilot programme. However, the same programme no longer improved learning outcomes when it was integrated into national education systems (Bold et al., 2018). Similarly, previous evaluations of NGO-supported pilot programmes in India demonstrated that grouping children by ability levels, and focusing on skills appropriate to those levels, had positive effects on learning outcomes. However, mainstreaming these changes into government schools led to important implementation challenges (Banerjee et al., 2016).

Our results aligned with these findings: children who participated in WCH's CWTL programme in Jordan (where CWTL replaced 40% of classroom time in the standard government curriculum) did not show larger learning gains than children in the comparison group after CWTL was integrated into the Jordanian national education system, and WUSC's remedial education programme did not show positive effects after being integrated into national schools in refugee camps. WCH's CWTL programme showed more positive effects on learning outcomes when targeting out-of-school children in Sudan, indicating that complex contextual factors can affect implementation and outcomes, including the specifics of the national education system (Brown et al., forthcoming).

Despite these findings, it remains important to scale education programmes by integrating innovations into national education systems. Banerjee and colleagues (2016) showed that grouping children by ability produced positive effects on learning outcomes, following several adaptations to programme design. After instituting careful, top-down support and monitoring to ensure that classrooms were reorganised around initial learning levels by government teachers, the programme achieved significant gains in language skills in the Indian states of Uttar Pradesh and Haryana (Banerjee et al., 2016). Similar adaptations may enable WCH, WUSC, and LWB to achieve gains in learning outcomes through their programmes. However, this would require several adaptations (informed by evidence) to ensure that the programmes remained both feasible to implement and effective in producing learning outcomes once integrated into national education systems.

Impacts on Psychosocial Outcomes

Mental health is a key component of the World Health Organisation's (WHO) definition of health and is important for enabling youth to reach their full potential in terms of both education and productivity (WHO, 2014). Children in conflict-affected settings often experience high rates of psychological distress. Evidence suggests that focused psychosocial interventions can reduce a range of symptoms and impairment associated with psychological distress amongst children in low-resource humanitarian settings (Purgato et al., 2018). However, despite their promise, a systematic review revealed a large evidence gap on the effects of education programmes on psychosocial outcomes (INEE, 2016).

Evidence is mixed on whether innovations in education have positive effects on psychosocial outcomes. We found mixed evidence on whether technology-in-education programmes had positive effects on psychosocial outcomes. In Jordan, we found that children attending CWTL schools demonstrated greater improvements in hope, as measured by the Child Hope Index. However, we did not find statistically significant differences between CWTL and comparison schools in changes in children's self-esteem, self-efficacy, or psychological well-being. In Lebanon, we found promising trends in children's self-esteem and psychological well-being that qualitative research attributed to CWTL, but the study did not have a comparison group. In Sudan, we found statistically significant effects of WCH's CWTL programme on children's self-esteem and psychological well-being, but we did not find statistically significant effects on children's hope (hope was not measured in Lebanon). The impact evaluations of WCH's CWTL programme in Jordan, Lebanon, and Sudan showed that digital, game-based learning programmes have the potential to improve psychosocial outcomes in some cases.

We also found mixed evidence on the effects of integrating LWB's Ideas Box into the DRC community centre. Qualitative evidence suggested that the Ideas Box may have contributed to improvements in psychosocial outcomes. However, a combination of propensity score matching and difference-in-difference analysis revealed that children in the comparison group showed larger, statistically significant improvements in psychosocial outcomes than children in the treatment group.[12] Both methods suffer from methodological limitations, so caution should be exercised when interpreting these results; more research is required to assess the impact of the Ideas Box on psychosocial outcomes. LWB did find positive effects of its Ideas Box on psychosocial outcomes in Colombia (CNC, 2018).

[12] Annex E presents the results of the impact evaluation of LWB's Ideas Box in the DRC community centre.

Impact evaluations of WUSC's remedial education programmes in Dadaab and Kakuma did not show statistically significant effects on either aspirations or resilience.

Labour Market Opportunities for Refugees

Employers in Rwanda with limited knowledge about hiring and recruiting refugees reported that they were less likely to hire refugees than Rwandans with identical characteristics. A triangulation of quantitative evidence and a literature review showed that refugees faced challenges in the labour market due to limited legal rights and limited employer knowledge about hiring and recruiting refugees. The literature review revealed that refugees still have limited legal rights in many countries, and returns to education decrease dramatically in contexts where refugees do not have the right to work, such as the Palestinian territories (Angrist, 1995).

In Rwanda, where refugees have the right to work, employers were 7 percentage points less likely to report that they would hire relatively well-educated refugees than host-population nationals with identical characteristics. However, providing information to employers about processes for hiring and recruiting refugees may improve refugees' labour market opportunities.[13] Analyses suggested that results in Rwanda were driven by employers who were less knowledgeable about recruiting refugees in the labour market; employers who reported having sufficient knowledge about hiring and recruiting refugees were just as likely to hire refugees as Rwandans with identical characteristics. This finding suggests that providing information to employers about processes for hiring and recruiting refugees may improve refugees' labour market opportunities.

A study by IDinsight (2019) shows that SNHU–Kepler graduates also reported a higher employment rate and were more likely to get a job immediately after graduation than a matched comparison group of students enrolled in other universities in Rwanda. In addition, more SNHU–Kepler graduates reported securing full-time formal employment than students in the matched comparison group (IDinsight, 2019).

[13] Annex B presents these results.

5. Discussion

The meta-evaluation findings focused on the scaling of innovations in education and their effects on educational and psychosocial outcomes in complex emergencies. They also provided some methodological lessons on the design and implementation of impact and process evaluations in protracted humanitarian crisis settings. Through adaptive programming and engagement with key stakeholders, each innovation team has designed and started to scale highly relevant education innovations that meet the needs of targeted participants. This approach is consistent with UNICEF's (2019) findings, which demonstrate that adaptation is the most common approach to scaling education programmes.

We initially adapted a conceptual framework (see the first column in Exhibit 21) that explained the journey to scale for innovations in education by distinguishing between the design phase (providing proof of concept), the scaling-up phase, and the scaling-out phase. Innovations go through the scaling-up phase to facilitate programme expansion within the same target group, while the scaling-out phase focuses on adapting the innovation to a different context in order to maintain effective programme implementation. The conceptual framework differs from the model developed by McClure and Gray (2015a), which focused less on where the innovation was used and more on the processes involved during scaling. Our framework also uses a different definition of scaling out: McClure and Gray (2015a) define scaling out as distilling complexity so that the solution is replicable and more easily adoptable, while we define scaling out as the replication of the programme in new contexts.

Our initial conceptual framework assumed that scaling up happens *before* scaling out. However, our empirical findings showed that most innovation teams started multiple pilot projects in different contexts, rather than scaling up in one context (before then scaling out). Our results showed that this decision was mostly driven by strategic considerations, donor location priorities, and uncertainty over future funding. Some innovation teams strategically chose to implement multiple pilots in new contexts to improve codification and generate evidence. However, other innovation teams were dependent on small grants, which created incentives to enter new contexts prioritised by donors, as well as uncertainty about funding going forward. This finding confirms that scaling can be a non-linear process (McClure & Gray, 2015a; Cooley & Linn, 2014; Results for Development & UNICEF, 2016).

We adapted the conceptual framework to reflect the role of education in emergencies and protracted crises, as well as the incentives and uncertainty about future funding created by the architecture for education in these settings. The adapted framework (see the second column in Exhibit 21) shows how incentives created

by donor location preferences, funding uncertainty, and the architecture for education in emergencies and protracted crises can increase the likelihood of starting multiple pilot projects in new contexts, and decrease the likelihood of expanding innovations in education within the same target group. The new framework differs from the original framework in its inclusion of these structural barriers (such as funding constraints), which create challenges for innovation teams in scaling innovations within the same target group.

The innovation teams highlighted that starting multiple pilot projects in new contexts was due, in part, to strategic considerations. Some teams operated in perpetual pilot mode (defined as the implementation of a pilot innovation in new contexts, driven by a new emergency or funding opportunity), which suggests that they were scaling out (Gray, 2019). Such strategies can be effective when they are accompanied by efforts to lay the groundwork for achieving sustainability—for example, by generating evidence, engaging with MoEs, and increasing the tangibility or codification of the programme. However, in some cases it was unclear whether the pilot in the new setting was scalable because some of those foundations for sustainability were missing (Gray, 2019).

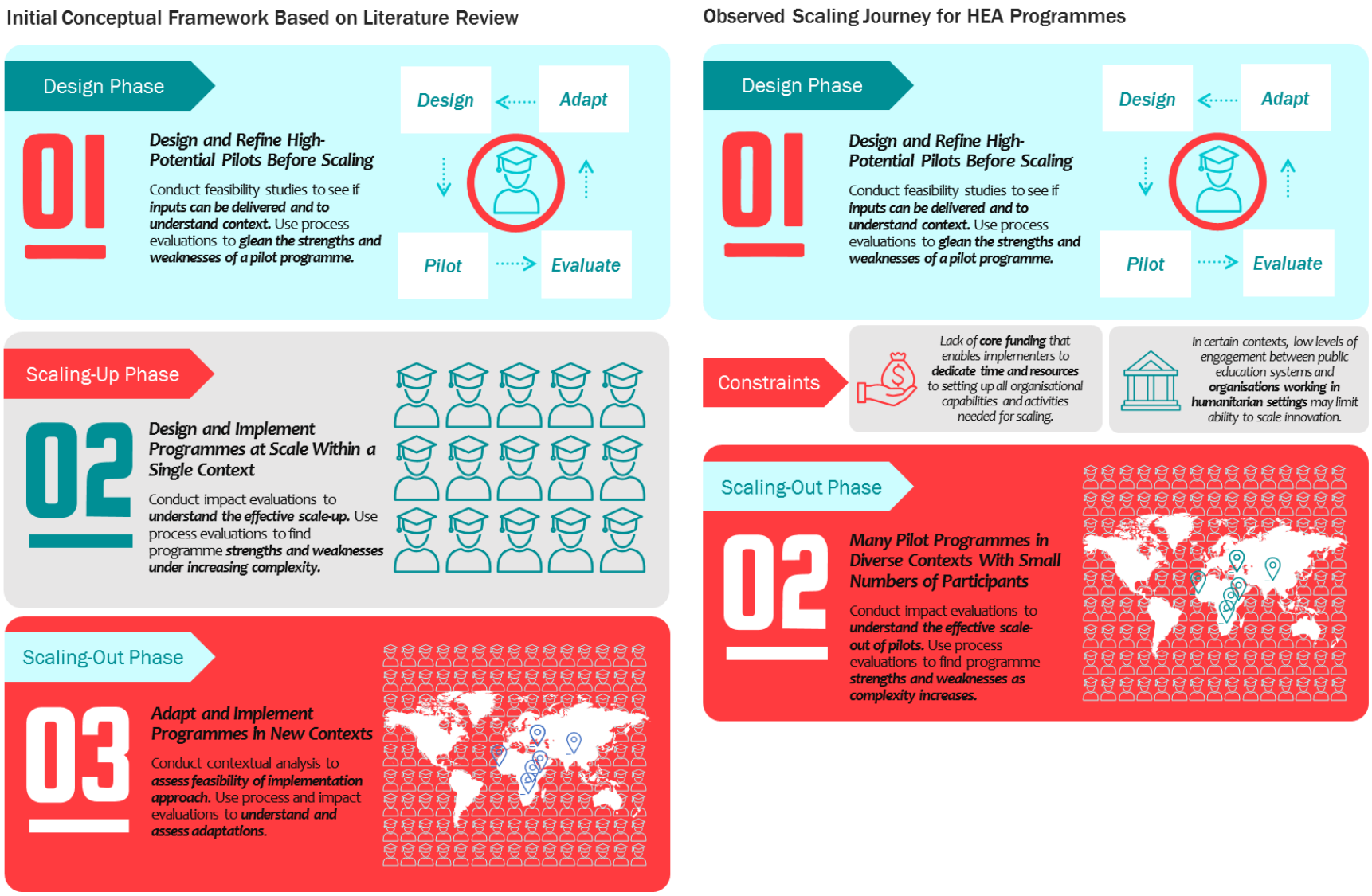
The discussion section is organised to reflect the phases in the revised conceptual framework: pilot, scale-up, and scale-out. We provide key findings on barriers to and facilitators of scaling, discuss the impact of innovations in education for each scaling phase, and present challenges that can compromise the methodological rigour of evaluations in humanitarian contexts. We conclude with policy recommendations based on these findings.

Pilot Phase

The new conceptual framework highlights the importance of designing and refining high-potential pilots before scaling. The results of our process and impact evaluations also demonstrate the importance of starting with a strong theory of change that is based on evidence about the impact of innovations in education, and that explicitly links programme activities to outputs and to intermediate and learning outcomes, as well as psychosocial outcomes.

We recognise that limited evidence is available about the impact of education innovations in humanitarian contexts. However, in the absence of evidence from humanitarian contexts, theories of change can be created using evidence from low- and middle-income countries (Bates & Glennerster, 2017). Although evidence from non-humanitarian contexts cannot always be credibly extrapolated to humanitarian contexts, evidence from one context can still be useful in other contexts. The complexity in humanitarian settings makes it particularly important to deliver evidence-based programme content that has proven successful in non-humanitarian settings (or ideally, in humanitarian settings, if the evidence is available), with relatively fixed core services supplemented by programme elements that can be adapted based on context. If a programme lacks a theoretical and structural basis for its core services, the complexity of a humanitarian setting makes it more difficult to deliver high-quality education to already vulnerable populations, particularly during scale-up or scale-out. Establishing *how* and *why* a programme and its theory of change differ in a humanitarian setting is critical, as innovations should be based on high-quality evidence of what works in education, specifically tailored to elements that are different in humanitarian settings. This highlights the importance of drawing on evidence from a wide range of settings to inform programme design and strategies, including development settings in low- and middle-income countries, in order to overcome the current evidence gaps in humanitarian contexts.

Exhibit 21. Initial Conceptual Framework Based on a Literature Review Versus an Observed Scaling Journey for HEA Programmes



Developing Education in Emergency Programmes Using Evidence-Based Education Practices

In line with the available literature, we propose a four-step process for using evidence from a wide range of settings and developing an evidence-based theory of change (Bates & Glennerster, 2017). Answering the following four questions will enable innovation teams to examine whether evidence from low- and middle-income countries can be credibly extrapolated to humanitarian contexts. This in turn can help implementers to create evidence-based theories of change in a programme's pilot phase.

1. What is the disaggregated theory behind the programme?
2. Do the local conditions hold for that theory to apply?
3. How strong is the evidence for the required general behavioural change?
4. What is the evidence that the implementation process can be carried out well?

This four-step process can help to determine the extent to which evidence from other contexts can be used to create (a) an evidence-based theory of change for a new innovation in a new context, (b) an evidence-based theory of change for implementing the programme at scale, and (c) an evidence-based theory of change for a different context. Of course, a programme that is underpinned by an evidence-based theory of change still may not achieve its objectives—for example, mechanisms that work in non-humanitarian contexts cannot always be credibly extrapolated to humanitarian contexts. In such cases, the wraparound services that constitute programme coordination and delivery can be adapted to suit humanitarian settings. Alternatively, a programme may not achieve its objectives because of implementation challenges. In such cases, it is critical to minimise the likelihood of implementation failure by redesigning the programme and monitoring the fidelity of programme implementation. The four-step process could also help implementers make decisions about adaptations to programme design. It is important to further refine a programme's theory of change during the pilot phase, especially when considering design adaptations. Investing in M&E and project management systems can help implementers to make these refinements.

Most teams used adaptations to respond to unexpected events or lessons from the field and develop processes for making these adaptations a priority. As a result, the five innovation teams were able to secure community support and generate demand for their programmes. However, the innovation teams did not incorporate human-centred design throughout the design and implementation of their programmes.

Overall, our research showed mixed evidence for positive impacts on learning outcomes and identified several implementation challenges that are specific to humanitarian settings, including the need for physical space, influxes of new students who are likely to have missed a lot of school, limited availability of quality teachers (trained teachers often do not speak the host country's language), and access issues. These challenges may all have contributed to the mixed evidence on positive impacts.

Cross-Sectoral Collaboration and Partnerships

Quantitative results identified opportunities to create synergies between education programmes and social protection or school feeding programmes. Although WUSC's remedial education programme did not have positive impacts on learning outcomes, on average, the analysis indicated that the programme could

have positive impacts on learning outcomes for girls who live in food-secure households and attend a minimum of 50 hours of remedial education per year. Cash transfers and school feeding programmes are likely to produce positive impacts on food security (Gilligan et al., 2013; Snilstveit et al., 2016), which could help education programmes to achieve positive impacts on learning outcomes.

Engaging With MoEs

Innovation teams' early engagement with MoEs during their programmes' pilot phases tended to focus on securing political buy-in and support for their interventions, rather than planning for long-term collaboration and scaling. Our literature review underscored the importance of engaging with governments and adapting education programmes for integration into national education systems in order to facilitate innovation scaling in education (de Hoop et al., 2018).

National governments can also be crucial actors in coordinating emergency responses and linking humanitarian and development programming (Overseas Development Institute, 2015). Engagement with MoEs is important not only because competing pressures often mean that education is a low priority in humanitarian settings, but also because engagement could help to mitigate challenges related to service integration, human capital, and finance. Although education budgets in humanitarian settings are often low (there is a global financing gap of at least $8.5 billion per year), domestic sources of funding—which include government spending and private household expenditures—are the largest source of funding for education in emergencies across all types of countries. The knowledge and skills of humanitarian organisations can complement the knowledge of government staff, who often do not have crisis or disaster management skills (Overseas Development Institute, 2015).

Achieving Impacts on Learning Outcomes When Integrating Programmes Into National Education Systems

The evaluation of WCH's CWTL programme in Sudan showed positive impacts on learning outcomes for out-of-school children. We hypothesise that technology-in-education programmes with a strong focus on pedagogical practices can contribute to improvements in learning outcomes for students in accelerated learning programmes who have limited previous access to education and start from a very low baseline. However, more research is needed to examine the external validity and transferability of this finding. The CWTL programme showed equal learning gains for in-school children using CWTL and for children following the standard government curriculum in Jordan, possibly because of the short period between baseline and endline data collection, or because positive impacts on learning outcomes are harder to achieve for students with higher baseline scores who are enrolled in public schools.

Scaling Out

Human-Centred Design

The innovation teams incorporated elements of human-centred design into their programmes but did not do so comprehensively throughout the design and implementation phases. Innovation teams are not alone in their inconsistent engagement with beneficiaries during the design phase; although practitioners seem to largely agree that user-centred design is important in humanitarian settings, and many guidelines and toolkits have been developed to facilitate community engagement in the programme design process, actual levels of beneficiary consultation vary considerably (Brown & Donini, 2014). Innovation teams

incorporated elements of human-centred design as they developed their programmes (e.g., adapting content to the context), but the interventions were predetermined rather than co-developed with the ultimate end users. Additionally, although teams understood the normative and operational value of maintaining open, two-way communication with beneficiary communities during implementation, Heller and colleagues (2011, p. 41) note that "constant communication" does not necessarily mean that a programme is truly accountable to beneficiaries, and that "dialogue does not necessary imply any fundamental change in the power imbalance or handing over of control or decision-making."

Scaling Up

Incentive to Scale Out Rather Than Up

Importantly, donor location preferences and a lack of longer term, flexible funding limited the ability of innovation teams to scale their education programmes within the same target group. Long-term funding channelled through government agencies and systems to support education is less prevalent in humanitarian settings because of a lower tolerance for financial risk and strict reporting requirements (Overseas Development Institute, 2015). These funding and coordination challenges in the architecture for education in protracted crisis settings create incentives to dedicate substantial resources to applying for small-scale funding. Innovation teams often faced uncertainty about future funding, which limited their ability to scale up in one context and achieve economies of scale. In addition, most innovation teams were dependent on many small grants from donors with different location preferences.

Methodological Research Lessons

Higher Number of Approvals Required for Data Collection in Humanitarian Contexts

The number of approvals required to conduct research in protracted humanitarian crisis settings may be even higher than in international development settings because approvals are needed from a larger number of key stakeholders, including local institutional review boards, relevant government institutions, and organisations that focus on refugee rights. For example, it was critical to obtain approval from the UNHCR to collect data in refugee camps in Kenya, Rwanda, and Jordan.

Importance of Flexibility in Planning for Data Collection

The nature of humanitarian contexts requires researchers to stay flexible when planning for data collection. Data collectors may only have limited access to insecure settings, especially after disruptive events. For example, we only had limited access to the Dadaab refugee camp because of insecurity. We mitigated that concern by relying on community mobilisers who resided in the camps to collect quantitative and qualitative data.

Importance of Training Individuals Who Reside in Insecure Settings and Are Trusted by the Community in Electronic Data Collection

Our experience during the HEA showed that for data collection purposes, it was important to provide data collection training to—and work with—individuals residing in insecure settings (such as refugee camps). Providing data collection training to these individuals can increase the quality of data because they are trusted by the community and familiar with the context.

6. Recommendations

Based on our meta-evaluation, we present recommendations that fall into two categories:

A. **Improving the overall architecture of education in emergencies, including the funding, coordination, and incentive structure.** These recommendations aim to help bilateral donors incentivise humanitarian organisations to refocus their efforts from pilot projects to programmes with solid business models that are more likely to have an impact on learning outcomes for a larger number of programme participants.

B. **Facilitating the scaling journey along the pilot, scaling-out, and scaling-up phases.** These recommendations aim to guide implementers in the use of evidence to scale innovations effectively in education.

Creating incentives that stimulate the scale-up of education programmes can encourage implementers to scale evidence-based education programmes. It could also enable a greater focus on using evidence and achieving improvements in learning outcomes in protracted crisis settings, including the implementation of Education Cannot Wait—a fund for education in emergencies that received $172 million in donors' contributions in its first year of operation. Currently, most education funding in emergencies focuses on small grants, with little emphasis on using evidence to stimulate improvements in learning outcomes. Education Cannot Wait aims to bridge the divide between humanitarian and development efforts, and by May 2019 it had invested $137 million in 19 crisis-affected countries. The fund has already contributed to increasing the share of humanitarian aid that goes to education in emergencies from 3.5% to 4% (UNICEF, 2018), and it recognises the importance of improving data collection, analysis, and tools. However, it places limited emphasis on creating incentives for moving effective education programmes to scale.

A. Recommendations to Improve the Architecture of Education in Emergencies

Our recommendations to improve the architecture for education in emergencies are intended to incentivise humanitarian organisations to refocus their efforts from pilot projects to programmes with solid management structures that can achieve improvements in learning outcomes for a large number of programme participants.

We recommend the following to policymakers and donors.

1. Allocate flexible funding that teams can use to solidify underlying management systems.

Allocating flexible funding can enable organisations to troubleshoot and solidify management systems before attempting to scale, which can guide implementers to ensure that they scale a comprehensive and sustainable programme. Currently, bilateral and multilateral donors focus on funding small grants for outward-facing aspects of programmes. However, this focus limits the ability of innovation teams to set up well-functioning business and administrative systems, even though these systems are necessary to scale programmes effectively. Allocating funding to plan and solidify systems would allow implementers to use resources freely to purposefully develop business systems that function in the long term. This includes building organisational capacities to scale (e.g., by ensuring that administrative systems are aligned with the scale-up) and supporting the business model and codification of the programme.

2. Focus education funding on larger scale education programmes for innovations that are in the final phase of the scaling process.

Funding larger scale programmes will enable implementers to focus on expanding education programmes in the same context in order to reach a larger number of programme participants, rather than starting multiple pilot projects in new contexts. The scaled-up programmes are likely to achieve greater cost-effectiveness thanks to economies of scale, in addition to increasing the likelihood of influencing learning outcomes by delivering refined content in a single context. However, it remains important to continue allocating funding to smaller scale innovations to limit the risk of disincentivising innovation.

3. Incentivise the use of evidence.

Incentivising the use of evidence can encourage implementers to use evidence from a wide range of settings, including development settings in low- and middle-income countries. Using evidence includes both ongoing use of monitoring, evaluation, and learning (MEL) systems, and theorising about causal pathways based on education programmes in non-humanitarian contexts. Such incentives can refocus humanitarian organisations' emphasis on evidence-based programming. For example, donors could provide funding to institutionalise the use of evidence through knowledge centres and evidence portals, and knowledge centres could produce reviews to inform decision-making about the funding of education programmes in humanitarian contexts (White, 2019). Ideally, areas of focus for these reviews should be based on regular meetings with donors, in which priority topics are agreed upon and emerging findings are discussed. Alternatively, evidence portals could enable evidence-based decision-making by providing recommendations informed by data (White, 2019). Evidence portals could also be complemented by guidelines and checklists that enable decision makers to use evidence without having to look at original research. These recommendations align with global best practices on the use of evidence, as outlined in the four waves of the evidence revolution (White, 2019).

4. Use performance-based financing.

Performance-based financing can motivate implementers to make adaptations that directly yield improvements in learning outcomes. Donors may, for example, decide to only fund education programmes that show improvements in learning outcomes based on rigorous evidence. In such cases, implementers may have more incentive to pilot innovations based on evidence of positive effects on learning outcomes.

Importantly, however, performance-based incentives only lead to improvements in programme design under certain conditions. Clist and Dercon (2014, p. 1) identify 12 principles to ensure that performance-based financing, or payment by results, "is used for the right things, in the right place, and in the right way." Examples of these principles include the use of agreed-upon measures (e.g., learning outcomes) and high-quality measurements (e.g., the Early Grade Reading Assessment for reading outcomes and the Early Grade Mathematics Assessment for mathematics outcomes). The 12 principles also suggest that implementers should have a relatively large degree of control over outcomes, which may make it more challenging to use performance-based incentives in humanitarian contexts. Nonetheless, performance-based incentives could stimulate the use of evidence with relatively low risk, if donors assume responsibility for a larger share of the risks (Clist & Dercon, 2014). For example, donors with widely diversified portfolios of projects could assume the risk of unexpected circumstances limiting the effects of an evidence-based intervention on learning outcomes (Clist & Dercon, 2014).

B. Recommendations to Facilitate the Scaling Journey

There are three phases in scaling innovations in education: the pilot phase, the scaling-out phase, and the scaling-up phase. First, we recommend designing and implementing programmes that have an evidence-based theory of change from the pilot phase onward (Pritchett et al., 2012). Second, it is important to adapt implementation in ways that allow programmes to be replicated more easily in different contexts during the scaling-out phase. Third, the scaling-up phase should focus on expanding an evidence-based programme within the same context. Each of the recommendations below aims to smooth the process for scaling innovations in education.

Pilot Phase

1. **Start building partnerships with MoE staff at all levels prior to implementation in order to learn about MoE structure, strategy, and priorities; how the government will sustain programming in the national education system; and options for financing.**

This recommendation aims to support implementers in developing partnerships with MoEs to facilitate scaling. Collaborative partnerships with MoEs are critical in humanitarian contexts to ensure the inclusion of all children and youth within safe, regulated, and accredited national education systems. Key components for success in developing relationships with MoEs include approaching the relationship as mutually supportive; engaging at every relevant level; presenting the MoE with a clear, results-based plan; including local experts; and being prepared for delays by building in time.

2. **Use a theory of change based on disaggregated theories from low- and middle-income countries and evidence from humanitarian contexts.**

Theories of change provide a structure for programme planning that makes programmes more likely to achieve outcomes by defining expected pathways, opportunities, and constraints (International Rescue Committee, 2019). Given the lack of literature on education programmes in humanitarian settings, programme theories of change should rely on literature about what works in education more broadly, particularly theories and practices that are well established (Bates & Glennerster, 2017). While evidence

from non-humanitarian contexts is not always applicable in humanitarian contexts, it can still be useful to strengthen the theory behind a programme's approach. The complexity in humanitarian settings makes it particularly important to deliver evidence-based content that has proven successful in non-humanitarian settings (or ideally, in humanitarian settings, if the evidence is available), with relatively fixed core services supplemented by programme elements that can be adapted based on context. Establishing a theory for *how* and *why* programme content differs in a humanitarian setting is critical for this approach, as innovations should be based on high-quality evidence of what works in education, specifically tailored to elements that are different in humanitarian settings. As discussed in Section 5, we recommend a four-step process to use evidence from other contexts and develop an evidence-based theory of change in the pilot phase (Bates & Glennerster, 2017), based on the following questions:

1. What is the disaggregated theory behind the programme?
2. Do the local conditions hold for that theory to apply?
3. How strong is the evidence for the required general behavioural change?
4. What is the evidence that the implementation process can be carried out well?

We provide one real-world example of how to apply this process, based on Evans (2017). This example focuses on a technology-based afterschool instruction programme in India that individualises content through technology and has had positive impacts on learning outcomes (Muralidharan, Singh, & Ganimian, 2019). The theory behind this programme is that classrooms contain students with very different abilities, but current teaching often does not reach all the student levels represented in a classroom. Students' learning outcomes could improve if technology enables students to learn at their own level, as in the technology-based afterschool instruction programme in India. Local conditions in humanitarian contexts suggest that this programme may be relevant because of the wide variation in learning levels in these contexts. However, refugees often do not speak the host country's language, which may mean that language adaptations are required to make the software contextually appropriate. An assessment of the strength of the evidence shows that an increasing number of studies suggest positive effects on learning outcomes from similar mechanisms that facilitate teaching at the students' instructional level (e.g., splitting classes by ability, providing teaching assistants to help the lowest performers; Duflo, Dupas, & Kremer, 2011; Kiesel & Duflo, 2015). However, implementation at scale may be challenging if teachers need to integrate the technology-based instruction programme in addition to the regular curriculum.

When applying this process, it is important to look beyond education outcomes. For example, our evidence on employer perceptions about hiring and recruiting refugees shows that refugees may experience challenges in the labour market in Rwanda if employers do not have sufficient knowledge about hiring and recruiting refugees. This suggests that providing information to employers about hiring and recruiting refugees may result in more opportunities for Rwandans in the labour market. Refugees' opportunities in the labour market are less promising in contexts outside Rwanda, where they are often not allowed to participate.

3. Adapt based on evidence.

The same four-step process can be used to help make decisions about adapting programme design. Adaptations should ideally help to achieve the ultimate objectives of education programmes—for example, improvements in learning outcomes. Assessing whether adaptations contribute to these objectives requires implementers to explicitly link them to intermediate and final outcomes in the

theory of change. The four questions presented above can guide implementers in examining whether adaptations that address barriers to effective programme implementation are ultimately likely to improve learning outcomes.

Ideally, adaptations that address barriers to effective programme implementation should be informed by strong formative and developmental evaluations in the pilot phase. These evaluations can guide implementers in examining the validity of the assumptions underlying the theory of change, leading to improvements in programme design (White, 2009; Pritchett et al., 2012).

4. Conduct rigorous research on the impact and implementation of programme components.

Just as education innovations should be developed based on evidence of what works to improve learning outcomes, programme components that are added or adapted require research to explore their effectiveness. Combining impact and process evaluations will yield concrete evidence of whether additions and adaptations are directly leading to improvements in learning outcomes, or whether resources could be better allocated to achieve the intended outcomes. Ideally, as Pawson and Tilley (2004, p. 2) note, rigorous research should be aligned with a realist evaluation approach in which evaluations aim to examine "[w]hat works, for whom, in what respects, to what extent, in what contexts, and how?"

5. Build and strengthen beneficiary feedback mechanisms to achieve accountability.

Innovation teams have achieved commendable support from the communities in which they operate, and they repeatedly emphasised the importance of community engagement. However, it was not always clear whether the teams provided an accessible feedback mechanism for beneficiaries and non-beneficiaries. As programmes scale, innovation teams should continue to build and strengthen beneficiary feedback mechanisms to enhance accountability, focusing not only on communication but also on full disclosure and "bottom-up" dialogue (Heller et al., 2011). Beneficiary feedback mechanisms can also guide adaptive programming based on community feedback.

Scaling-Out Phase

1. Focus on scaling out education innovations when local conditions and the delivery model in the new context are similar to the local conditions and delivery models in contexts where evidence on programme effectiveness was gathered.

Most innovation teams either changed the delivery model when scaling out to new contexts or scaled out to contexts with different local conditions. As a result, it remains unclear whether evidence on implementation challenges and programme effects that was gathered in the initial context can be credibly extrapolated to the new context. Furthermore, it is often unclear whether an initial theory of change applies in new local conditions. Implementers could generate new evidence on programme effectiveness in the new context, but the costs of this are often prohibitive. In addition, there will be considerable uncertainty about the effectiveness of the scaled-out programme until the impact evaluation has been finalised. It is also important to focus on potential cost savings in the scaling-out process, and McClure and Gray (2015a) argue that customising solutions to meet local needs may deplete already scarce resources.

2. Prioritise early development of management systems capable of supporting operations in multiple contexts.

Early prioritisation of the logistical and management aspects of a programme—including streamlining administrative processes, financial tracking, and partnership procedures—not only mitigates later challenges that result from a lack of organisation, but also creates a template for scaling pilot programmes in an efficient and effective manner. This recommendation aligns with UNICEF's suggestion that innovations establish a "comprehensive packaged and globally deployable solution set that distinguish[es] core repeatable content and processes from components that must be localised, and entry points for customisation" (2019, p. 10). Because funding usually prioritises the outward-facing elements of a programme, teams should include the cost of solidifying the less visible components of their programmes in their budget proposals. Cost analyses that demonstrate how solidifying procedures at the outset can achieve cost-effectiveness could help to demonstrate the value of investing in systems at the pilot stage. Unrestricted funding for the creation of business systems to move to scale will also help to prioritise the early development of management systems capable of supporting operations in multiple contexts.

Scaling-Up Phase

1. Do not scale up unless evidence shows that the programme contributes to improvements in learning outcomes.

We only recommend scaling education programmes in humanitarian settings when evidence suggests that the innovation can effectively contribute to achieving key programme objectives. Our evidence suggests that several innovations will not result in improvements in learning outcomes without significant adaptations to programme design. WCH's CWTL programme did not show differential gains in learning outcomes between students enrolled in schools that implemented CWTL 40% of the time and children following the standard government curriculum in Jordan. WUSC's remedial education programme in Kenya also did not demonstrate positive effects on learning outcomes, perhaps due to significant challenges in programme implementation. We recommend that implementers make evidence-based adaptations to programme design before they consider moving these programmes to scale.

In contrast, WCH's CWTL programme in Sudan showed large and statistically significant effects on learning and psychosocial outcomes amongst out-of-school children. This finding suggests that technology solutions can improve learning outcomes for out-of-school children in some contexts. Although it is critical to assess programme impacts at scale, these results also suggest that it may be feasible to scale the CWTL programme in Sudan and achieve improvements in learning outcomes. However, the different outcomes in Jordan and Sudan indicate that complex contextual factors can affect implementation and outcomes, including the specifics of the national education system.

2. Use a human-centred design approach to inform programme design and scaling decisions.

UNICEF (2019) identifies human-centred design as a promising approach for considering feasibility, viability, and desirability when making decisions about scaling. Innovation teams consistently referred to challenges in making decisions about how to scale their programmes, which suggests that adopting a stronger human-centred design approach could potentially help future innovators to make more informed

decisions about scaling. However, there is a dearth of rigorous evidence on how incorporating user-centred design ultimately affects outcomes (Brown & Donini, 2014). It is critical to design and implement more research into how incorporating human-centred design principles can contribute to learning and psychosocial outcomes in humanitarian settings. Evidence from other sectors shows that community-based practices do not always deliver positive effects (Mansuri & Rao, 2013; White, 2019), which highlights the importance of assessing the effectiveness of human-centred design approaches.

References

Agha, S. (2016). *Advancing adolescents: Evidence on the impact of psychosocial support for Syrian refugee and Jordanian adolescents.* Retrieved from https://www.mercycorps.org/sites/default/files/Advancing_Adolescents%20Report_FINAL_ONLINE.pdf

American Institutes for Research. (2016a). *Evaluation of a UNICEF cash transfer for displaced Syrian children in Lebanon: Baseline report.* Washington, DC: Author.

American Institutes for Research. (2016b). *Early grade reading in Latin America and the Caribbean: A systematic review. LAC Reads Capacity Program*. Retrieved from https://lacreads.org/sites/default/files/lrcp-systematic-rvw-rpt-508.pdf

Angrist, J. D. (1995). The economic returns to schooling in the West Bank and Gaza Strip. *The American Economic Review*, 1065–1087.

Atkins, S., Lewin, S., Smith, H., Engler, M., Fretheim, A., & Volmink, J. (2008). Conducting a meta-ethnography of qualitative literature: Lessons learnt. *BMC Medical Research Methodology, 8*(21).

Baird, S., Ferreira, F., Özler, B., & Woolcock, M. (2014). Conditional, unconditional and everything in between: A systematic review of the effects of cash transfer programs on schooling outcomes. *Journal of Development Effectiveness, 6*(1), 1–43. doi:10.1080/19439342.2014.890362

Banerjee, A., Duflo, E., Goldberg, N., Karlan, D., Osei, R., Pariente, W., . . . Udry, C. (2015). A multifaceted program causes lasting progress for the very poor: Evidence from six countries. *Science*, *348*(6236), 1260799.

Banerjee, A. V., Cole, S., Duflo, E., & Linden, L. (2007). Remedying education: Evidence from two randomized experiments in India. *Quarterly Journal of Economics*, *122*(3), 1235–1264.

Banerjee, A., Banerji, R., Berry, J., Duflo, E., Kannan, H., Mukerji, S., . . . Walton, M. (2017). From proof of concept to scalable policies: Challenges and solutions, with an application. *Journal of Economic Perspectives*, *31*(4), 73–102.

Barrera-Osorio, F., & Linden, L. L. (2009). *The use and misuse of computers in education: Evidence from a randomized controlled trial of a language arts program.* Policy Research Working Paper Series – The World Bank (March).

Bates, M. A., & Glennerster, R. (2017). The generalizability puzzle. *Stanford Social Innovation Review.* Retrieved from https://ssir.org/articles/entry/the_generalizability_puzzle

Benson, C. 2009. Designing effective schooling in multilingual contexts. In A. K. Mohanty, R. Phillipson, & T. Skutnabb-Kangas (Eds.), *Social justice through multilingual education* (pp. 63–81). Clevedon, UK: Multilingual Matters.

Bennell, P. (2004). *Teacher motivation and incentives in sub-Saharan Africa and Asia*. Brighton: Knowledge and Skills for Development. Retrieved from https://pdfs.semanticscholar.org/0009/cc7e647c18899afbbf6b19a3362591fd8b32.pdf

Bessant, J., Ramalingam, B., Rush, H., Marshall, N., Hoffman, K., & Gray, B. (2014). *Innovation management, innovation ecosystems and humanitarian innovation.* Retrieved from https://assets.publishing.service.gov.uk/media/57a089eb40f0b652dd000480/Humanitarian-Innovation-Ecosystem-research-litrev.pdf

Betts, A. & Bloom, L. (2014). Humanitarian innovation: The state of the art. *OCHA Policy and Studies Series: Occasional Policy Paper.* Retrieved from https://www.unocha.org/sites/unocha/files/Humanitarian%20Innovation%20The%20State%20of%20the%20Art_0.pdf

Bold, T., Kimenyi, M., Mwabu, G., Ng'ang'a, A., & Sandefur, J. (2018). Experimental evidence on scaling up education reforms in Kenya. *Journal of Public Economics, 168*, 1–20.

Borenstein, M., Hedges, L. V., Higgins, J. P. T., & Rothstein, H. (2009). *Introduction to meta-analysis.* Chichester, UK: Wiley.

Brody, C., de Hoop, T., Vojtkova, M., Warnock, R., Dunbar, M., Murthy, P., & Dworkin, S. L. (2015). Economic self-help group programs for women's empowerment: A systematic review. *Campbell Systematic Reviews*, *19.*

Brody, C., de Hoop, T., Vojtkova, M., Warnock, R., Dunbar, M., Murthy, P., & Dworkin, S. L. (2017). Can self-help group programs improve women's empowerment? A systematic review. *Journal of Development Effectiveness, 9*(1), 15–40.

Brookings Institution. (2008). *Moving beyond rhetoric: Participation with populations displaced by conflict or natural disasters*. Washington, DC: Brookings Institution and University of Bern Project on Internal Displacement.

Brown, D., & Donini, A. (2014). *Rhetoric or reality? Putting affected people at the centre of humanitarian action*. London: ALNAP/ODI.

Bulman, G., & Fairlie, R. W. (2016). *Technology and education: Computers, software, and the internet.* (No. 9432). IZA Discussion Papers.

Burde, D., Guven, O., Kelcey, J., Lahmann, H., & Al-Abbadi, K. (2015). *What works to promote children's education access, quality of learning and well-being in crisis-affected contexts: Education rigorous literature review*. London, UK: Department for International Development (DFID). Retrieved from https://www.gov.uk/government/uploads/system/uploads/attachment_data/file/470773/Education-emergencies-rigorous-review2.pdf

Cabezas, V., Cuesta, J. I., & Gallego, F. A. (2011). *Effects of short-term tutoring on cognitive and non-cognitive skills: Evidence from a randomized evaluation in Chile*. Unpublished working paper.

Caliendo, M., & Kopeinig, S. (2008). Some practical guidance for the implementation of propensity score matching. *Journal of Economic Surveys, 22*(1), 31–72.

Cameron, L., & Shah, M. (2017). *Scaling up sanitation: Evidence from an RCT in Indonesia*. IZA Discussion Paper No. 10619. Retrieved from http://ftp.iza.org/dp10619.pdf

Chapelier, C., & Shah, A. (2013) *Improving communication between humanitarian aid agencies and crisis-affected people: Lessons from the infoasaid project*. HPN Network Paper 74. London: ODI/HPN.

Chau, K., Seck, A. T., Chandra-Mouli, V., & Svanemyr, J. (2016). Scaling up sexuality education in Senegal: Integrating family life education into the national curriculum. *Sex Education*, *16*(5), 503–519. doi:10.1080/14681811.2015.1123148

Chaffin, J. (2016). *Norwegian Refugee Council Jordan youth programme evaluation*. Retrieved from https://www.nrc.no/resources/evaluations/nrc-jordan-youth-programme-evaluation/

Chinen, M., Coombes, A., De Hoop, T., & Elmeski, M. (2016). *Evaluation of the transformative potential of positive gender socialization in education for peace building.* Retrieved from http://www.air.org/sites/default/files/downloads/report/Evaluation-Transformative-Potential-Positive-Gender-Socialization-Peace-Building-May-2016.pdf

Chinen, M., Coombes, A., de Hoop, T., Elmeski, M., & Castro Zazur, R. (2017). Can teacher training and SMS messages change gender norms? Evidence from Northern Uganda. *Journal on Education in Emergencies*, *3*(1), 44–78.

Cilliers, J., & Taylor, S. (2017). *How to re-skill a workforce? Experimental evidence of in-service training and coaching.* Working Paper.

Clist, P., & Dercon, S. (2014.) 12 Principles for payment by results (PbR) in international development. Norwich: University of East Anglia. Retrieved from https://assets.publishing.service.gov.uk/media/57a089d2e5274a27b20002a5/clist-dercon-PbR.pdf.

CNC. (2018.) Evaluación de impacto del proyecto bibliotecas públicas móviles. Retrieved from https://www.bibliosansfrontieres.org/wp-content/uploads/2019/08/Informe-de-resultados-Evaluacio%CC%81n-de-impacto-BPM.pdf

Cooley, L., & Linn, J. F. (2014). *Taking innovations to scale: Methods, applications, and lessons.* Retrieved from https://www.usaid.gov/sites/default/files/documents/1865/v5web_R4D_MSI-BrookingsSynthPaper0914-3.pdf

Crea, T. M., & McFarland, M. (2015). Higher education for refugees: Lessons from a 4-year pilot project. *International Review of Education, 61*(2), 235–245.

Dahya, N. (2016). *Education in conflict and crisis: How can technology make a difference? A landscape review*. Retrieved from http://www.educationinnovations.org/research-and-evidence/education-conflict-and-crisis-how-can-technology-make-difference-landscape

De Hoop, T., Brudevold-Newman, A., Rothbard, V., Todd, A., Rai, N., & Kiggundu, G. (2018). *Moving from pilot to scale in protracted humanitarian crisis settings: An evidence-informed conceptual framework for the Humanitarian Education Accelerator*. Washington, DC: American Institutes for Research.

Duflo, E., Dupas, P., & Kremer, M. (2011). Peer effects, teacher incentives, and the impact of tracking: Evidence from a randomized evaluation in Kenya. *American Economic Review, 101*(5), 1739–1774.

Dupas, P. (2011). Do teenagers respond to HIV risk information? Evidence from a field experiment in Kenya. *American Economic Journal: Applied Economics*, *3*, 1–34. doi:10.1257/app.3.1.1

Easton-Calabria, E., & Omata, N. (2016). *Micro-finance in refugee contexts: Current scholarship and research gaps*. Refugee Studies Centre Working Paper Series No. 116.

El Daoi, R. (2019.) *Many refugee children must work to feed their families*. Norwegian Refugee Council report. Reliefweb. Retrieved from https://reliefweb.int/report/lebanon/many-refugee-children-must-work-feed-their-families

Elrha. (2018). *Too tough to scale? Challenges to scaling innovation in the humanitarian sector*. London: Author.

Epstein, A. (2014). *Tusome baseline study: Baseline study for future evidence performance evaluation.* Washington, DC: USAID.

Evans, D. (2017). *A framework for taking evidence from one location to another*. Retrieved from http://blogs.worldbank.org/impactevaluations/framework-taking-evidence-one-location-another

Evans, D. K., & Popova, A. (2015). *What really works to improve learning in developing countries?* Working Paper. The World Bank Group. Retrieved from http://documents.worldbank.org/curated/en/516191468172488103/pdf/WPS7203.pdf

Evans, D., & Yuan, F. (2019). *Equivalent years of schooling: A metric to communicate learning gains in concrete terms.* World Bank Policy Research Working Paper 8752. Retrieved from http://documents.worldbank.org/curated/en/123371550594320297/pdf/WPS8752.pdf

Fehr, E., & Gächter, S. (2000). Fairness and Retaliation: The Economics of Reciprocity." *Journal of Economic Perspectives,* 14(3), 159-181.

Gilligan, D. O., Margolies A., Quinones E., & Roy, S. (2013). *Impact evaluation of cash and food transfers at early childhood development centers in Karamoja, Uganda*. Retrieved from https://www.unicef.org/evaldatabase/files/Impact_Evaluation_of_Cash_and_Food_Transfers_at_Early_Childhood_Development_Centers_in_Karamoja_-_Uganda.pdf

Glewwe, P., & Muralidharan, K. (2015). *Improving school education outcomes in developing countries*. Working Paper 15/001, RISE.

Global Innovation Fund. (n.d.) *The potential of scale*. Retrieved from https://www.globalinnovation.fund/what-we-do/how-we-scale/

Gove, A., Poole, M., & Piper, B. (2017) Designing for scale: Reflections on rolling out reading improvement in Kenya and Liberia. *New Directions for Child and Adolescent Development, 155*, 77–95.

Gray, I. (2019). *Scaling Innovations in Education in Emergencies – learnings from the Humanitarian Education Accelerator.*

Grootenhuis, F., & Calo, M. (2016). *Testing new ground: A case study on multisector cash interventions in Mangaize Refugee Camp, Niger*. Retrieved from http://www.alnap.org/pool/files/calp-multisector-cash-interventions-in-niger-casestudy.pdf

Gutiérrez, E., & Rodrigo, R. (2014). Closing the achievement gap in mathematics: Evidence from a remedial program in Mexico City. *Latin American Economic Review*, *23(*14), 1–30.

Hartmann, A., & Linn, J. F. (2007). Scaling up: A path to effective development. *Twenty twenty (2020) focus brief on the world's poor and hungry people/International Food Policy Research Institute (IFPRI)*.

Heller, O., Költzow, S., & Vasudevan, R. (2011). *Formal systems of constant dialogue with host societies in humanitarian projects.* Graduate Institute of International Development Studies. Geneva, Switzerland.

Humphreys, M., Sanchez de la Sierra, R., & Van der Windt, P. (2014). *Social and economic impacts of Tuungane: Final report on the effects of a community-driven reconstruction programme in the Democratic Republic of Congo*. 3ie Impact Evaluation Report 7. New Delhi: International Initiative for Impact Evaluation.

IDEO. (2015). *Human-centered design toolkit: 2nd edition*. Retrieved from https://hcd-connect-production.s3.amazonaws.com/toolkit/en/download/ideo_hcd_toolkit_final_cc_superlr.pdf

IDinsight. (2016). *Kepler Kiziba qualitative study of gender, curriculum, and employment opportunities.* Kigali, Rwanda: IDinsight.

IDinsight. (2019). *SNHU-Kepler impact evaluation: 2019 endline results.* Kigali, Rwanda: IDinsight.

INEE. (2016.) *Psychosocial support and social and emotional learning for children and youth in emergency settings*. Background paper. Retrieved from https://inee.org/system/files/resources/INEE_PSS-SEL_Background_Paper_ENG_v5.3.pdf

International Rescue Committee. (2019.) *The outcomes and evidence framework.* Retrieved from http://oef.rescue.org/#/home?_k=oglihd.

Jones, A. (Forthcoming). *Journey to scale: Value for money of education in emergency programmes.* War Child Holland. Amsterdam, Netherlands.

Kiesel, J., & Duflo, A. (2015). *Targeted lessons to improve basic skills.* Retrieved from https://www.poverty-action.org/sites/default/files/publications/TCAI_Final%20Results_040115.pdf

Lacey, C., Cooper, B., & Torrance, H. (1993). Evaluating the Andhra Pradesh primary education project: Problems of design and analysis. *British Educational Research Journal*, *19*(5), 535–554.

Lakshminarayana, R., Eble, A., Bhakta, P., Frost, C., Boone, P., Elbourne, D., . . . Mann, V. (2013). The support to rural India's public education system (STRIPES) trial: A cluster randomised controlled trial of supplementary teaching, learning material and material support. *PLOS One, 8*(7).

Larson, R. S. L., Dearing, R., & Backer, T. (2017). Strategies to scale up social programs. Retrieved from https://comartsci.msu.edu/sites/default/files/documents/research/Larson-Dearing-Backer-2017.pdf

Li, T. (2012). Remedial education with incentivized peer-tutoring: Evidence from migrant children schools in China. *Developing Economies*, *50*(4), 334–350. doi:10.1111/j.1746-1049.2012.00178.x

Lucas, A. M., McEwan, P. J., Ngware, M., & Oketch, M. (2014). Improving early-grade literacy In East Africa: Experimental evidence from Kenya And Uganda. *Journal of Policy Analysis and Management*, *33*(4), 950–976.

MacLaren, D. (2010). Tertiary education for refugees: A case study from the Thai-Burma border. *Refuge*, *27*(2), 103–110.

Mansuri, G., & Rao, V. (2013). *Localizing development: Does participation work?* The World Bank Group: Washington, DC.

McClure, D., & Gray, I. (2015a). Scaling: Innovation's missing middle. Submitted for the *Transformation Through Innovation Theme for the World Humanitarian: A Landscape Review, 65*.

McClure, D., & Gray, I. (2015b). *Managing the journey to scale up innovation: In the humanitarian and development sector*. Submitted to the Transformation Through Innovation Theme for the World Humanitarian Summit.

McClure, D., & Gray, I. (2015c). *Engineering scale up in humanitarian innovations' missing middle*. doi:10.1109/GHTC.2015.7343963

McEwan, P. J. (2015). Improving learning in primary schools of developing countries. *Review of Educational Research, 85*(3), 353–394. doi:10.3102/0034654314553127

Melaville, A., Jacobson, R., & Blank, M. J. (2011). Scaling up school and community partnerships: The community schools strategy. *Washington, DC: Coalition for Community Schools*.

Muralidharan, K., Singh, A., & Ganimian, A. J. (2019). Disrupting education? Experimental evidence on technology-aided instruction in India. *American Economic Review*, *109*(4), 1426–1460.

Nakamura, P., de Hoop, T., & Udayakumar Holla, C. (2018). Language and the learning crisis: Evidence of transfer threshold mechanisms in multilingual reading in South India. *Journal of Development Studies*, 55(11), 2287–2305.

Nicolai, N., & Hine, S. (2015). *Investment for education in emergencies: A review of evidence.* Retrieved from https://www.usaid.gov/sites/default/files/documents/1865/Investment_Emergencies_Education.pdf

Obrecht, A., & Warner, A. (2016.) *More than just luck: Innovation in humanitarian action*. Humanitarian innovation fund. Retrieved from https://www.elrha.org/wp-content/uploads/2017/03/hif-alnap-more-than-just-luck-2016.pdf.

Obrecht, A., Warner, A. T., & Dillon, N. (2017). *Evaluating humanitarian innovation.* HIF/ALNAP Working Paper. London, UK: ODI/ALNAP.

Ouane, A., & Glanz, C. (2010*). Why and how Africa should invest in African languages and multilingual education: An evidence and practice-based policy advocacy brief.* UNESCO Institute for Lifelong Learning. Retrieved from https://unesdoc.unesco.org/ark:/48223/pf0000188642

Overseas Development Institute. (2014). *Doing development differently: Who we are, what we're doing, what we're learning.* Retrieved from https://www.odi.org/sites/odi.org.uk/files/resource-documents/11199.pdf?_sm_au_=iMVrLr1tDQ5tW0jH

Overseas Development Institute. (2015). *How cash transfers can transform humanitarian aid. Report of the High Panel on Humanitarian Cash Transfers.* London: Author. Retrieved from https://www.odi.org/publications/9876-cash-transfers-humanitarian-vouchers-aid-emergencies

Panirsilvam, S. (2017). *Literature review: Scaling up of education innovation and their impact on students learning outcomes*. Retrieved from https://doi.org/10.13140/RG.2.2.25120.28169

Pawson, R. & Tilley, N. (2004). Realist evaluation. *British Cabinet Office.* Retrieved from https://pdfs.semanticscholar.org/011c/82d379c6525e8402bd0bd39a780b94297a78.pdf

Pick de Weiss, S., Andrade-Palos, P., Townsend, J., & Givaudan, M. (1994). Evaluation of a sexual education program on knowledge, sexual behavior, and contraception among adolescents. *Salud Mental, 17*(1), 25–31.

Pingel, F. (2010). The power of curriculum. In K. M. Cahill (Eds.), *Even in chaos: Education in times of emergency* (pp. 109–135). New York, NY: Fordham University Press.

Piper, B., Destefano, J., Kinyanjui, E. M., & Ong'ele, S. (2018). Scaling up successfully: Lessons from Kenya's Tusome national literacy program. *Journal of Education Change 19*(3), 293. doi:10.1007/s10833-018-9325-4.

Piper, B., & Korda, M. (2011). *EGRA Plus: Liberia*. Program evaluation report. RTI International.

Piper, B., & Mugenda, A. (2013). *The Primary Math and Reading (PRIMR) Initiative midterm impact evaluation.* Washington, DC: USAID.

Piper, B., Zuilkowski, S. S., & Mugenda, A. (2014). Improving reading outcomes in Kenya: First-year effects of the PRIMR Initiative. *International Journal of Educational Development*, *37*, 11–21.

Popova, A., Evans, D. K., & Arancibia, V. (2016). *Training teachers on the job: What works and how to measure it*. The World Bank.

Pritchett, L., Samji, S., & Hammer, J. (2012). *It's all about MeE: Using structured experiential learning ("e") to crawl the design space.* Center for Global Development: Washington, DC.

Purgato, M., Gross, A. L., Betancourt, T., Bolton, P., Bonetto, C., Gastaldon, C., . . . Barbui, C. (2018). Focused psychosocial interventions for children in low-resource humanitarian settings: A systematic review and individual participant data meta-analysis. *Lancet Global Health 6*(4), e390–e400.

Puri, J., Aladysheva, A., Iversen, V., Ghorpade, Y., & Brück, T. (2015). *What methods may be used in impact evaluations of humanitarian assistance?* Bonn, Germany: Institute for the Study of Labor.

Purnell, S., & Kengkunchorn, A. (2008). *Taking learning further: A research paper on refugee access to higher education.* Mae Sot, Thailand: Simon Purnell.

Ramalingam, B., Rush, H., Bessant, J., Marshall, N., Gray, B., Hoffman, K., . . . Warren, K. (2015). *Humanitarian innovation ecosystem research project: Final report.*

Refugee Studies Centre. (2005). *Lessons learned report: The transnationalisation of care Sahrawi refugee children in a Spanish host program*. Retrieved from http://www.southampton.ac.uk/studentadmin/about_saa/saa_business_review_projects/Completed_Projects/docs/Microsoft Word—Health Sciences Integration Project—Lessons Learned report.pdf

Results for Development & UNICEF (2016). *Journeys to scale: Accompanying the finalists of the innovations in education initiative.* Retrieved from https://www.r4d.org/resources/journeys-scale-accompanying-finalists-innovations-education-initiative/

Rose, P., & Greeley, M. (2006). *Education in fragile states: Capturing lessons and identifying good practice.* DAC Fragile States Group.

Ring, H. R., & West, A. R. (2015). Teacher retention in refugee and emergency settings: The state of the literature. *International Education Journal: Comparative Perspectives*, *14*(3), 106–121.

Snilstveit, B., Stevenson, J., Menon, R., Phillips, D., Gallagher, E., Geleen, M., . . . Jimenez, E. (2016). *The impact of education programmes on learning and school participation in low- and middle-income countries*. Retrieved from http://www.3ieimpact.org/media/filer_public/2016/09/20/srs7-education-report.pdf

Stone, R., de Hoop, T., Coombes, A., Goett, M., Morey, M., & Kamto, K. (2018). *What works to improve early grade literacy in Latin America and the Caribbean: 2016 update of a systematic review*. USAID.

Stubbé, H., Badri, A., Telford, R., Van der Hulst, A., & Van Joolingen, W. (2016). E-learning Sudan, formal learning for out-of-school children. *Electronic Journal of e-Learning, 14*(2), 136–149.

Tinajero, A. R. (2010). *Scaling-up early child development in Cuba: Cuba's Educate Your Child Program: Strategies and lessons from the expansion process* (Wolfensohn Center for Development No. 16).

Tull, K. (2018). *School feeding interventions in humanitarian responses*. University of Leeds Nuffield Centre for International Health and Development Róisín Plunkett, Education Development Trust. https://assets.publishing.service.gov.uk/media/5be5b217ed915d6a0d6f6fcf/360_School_Feeding_Interventions_in_Humanitarian_Responses.pdf

Tulloch, C. (2016). *Cost efficiency: unconditional cash transfers.* Retrieved from https://www.rescue.org/report/cost-efficiency-unconditional-cash-transfers

UNHCR. (2017). *Turn the tide: Refugee education in crisis*. Retrieved from https://www.unhcr.org/5b852f8e4.pdf

UNHCR Jordan. (2018.) *Fact sheet: Jordan*. Reliefweb. Retrieved from https://reliefweb.int/sites/reliefweb.int/files/resources/FactSheetJordanFebruary2018-FINAL_0.pdf

UNHCR Rwanda. (2019.) *Monthly population statistics*. Reliefweb. Retrieved from https://reliefweb.int/sites/reliefweb.int/files/resources/70668.pdf

UNICEF. (2018). *Education cannot wait: A fund for education in emergencies: Results report,* April 2017–March 2018. Retrieved from https://reliefweb.int/report/world/education-cannot-wait-results-report-april-2017-march-2018

UNICEF. (2019). *Scaling Innovation for every child: Strategies, models and critical success factors that emerged from 1,600+ days of scaling innovations for children across 90 countries*. Author.

UNICEF Bangladesh. (2019.) *Beyond survival: Rohingya refugee children in Bangladesh want to learn*. UNICEF Advocacy Alert. Reliefweb. Retrieved from https://reliefweb.int/sites/reliefweb.int/files/resources/UN0336024.pdf

UNICEF Kenya. (2019.) *UNICEF Kenya humanitarian situation report*. Reliefweb. Retrieved from https://reliefweb.int/sites/reliefweb.int/files/resources/UNICEF%20Kenya%20Humanitarian%20Situation%20Report%20mid-year%202019.pdf

United Nations Evaluation Group. (2008). *UNEG Code of Conduct for Evaluation in the UN System.* Retrieved from http://www.unevaluation.org/document/detail/100

Uwezo. (2015). *Trainers & district coordinators manual.* Uwezo Kenya. Retrieved from https://palnetwork.org/wp-content/uploads/2016/2016Tools/003_2015_KE_TrainersManual.pdf

Valters, C., Cummings, C. & Nixon, H. (2016). *Putting learning at the centre: adaptive development programming in practice*. Overseas Development Institute. Retrieved from https://www.odi.org/sites/odi.org.uk/files/resource-documents/10401.pdf

Vogel, I. (2012). *Review of the use of 'Theory of Change' in International development*. DFID Research Paper. Department for International Development.

Waddington, H., White, H., Snilstveit, B., Hombrados, J. G., Vojtkova, M., Davies, P., . . . Valentine, J. C. (2012). How to do a good systematic review of effects in international development: A tool kit. *Journal of Development Effectiveness, 4*(3), 359–387.

White, H. (2009). Theory-based impact evaluation: Principles and practice. *Journal of Development Effectiveness*, *1*(3), 271–284.

White, H. (2019). The twenty-first century experimenting society: The four waves of the evidence revolution. *Palgrave Communications, 5*(47).

Winthrop, R., & Kirk, J. (2008). Learning for a bright future: Schooling, armed conflict, and children's well-being. *Comparative Education Review, 52*(4), 639–661. Retrieved from http://www.jstor.org/stable/10.1086/591301

World Bank. (2018). *The world development report 2018: LEARNING to realize education's promise*. Retrieved from https://www.worldbank.org/en/publication/wdr2018

World Health Organization (WHO). (2014). *Mental health: A state of wellbeing*. Retrieved from https://www.who.int/features/factfiles/mental_health/en/

Annex A. Division of Roles in Conducting the Evaluation

Exhibit A-1. Division of Roles Between AIR and the Innovation Teams in Conducting the Evaluation

Innovation Team (Programme)	Evaluation	Evaluation Design	Data Collection	Data Analysis
War Child Holland (WCH), Can't Wait to Learn (CWTL) programme	Impact evaluation in Jordan	Lead: AIR Support: WCH	Lead: Local data collection firm quality assurance Enumerator training: AIR/WCH	Lead: AIR Support: WCH
	Practice-oriented evaluation in Lebanon and impact evaluation in Sudan	Lead: WCH Support: AIR/Netherlands Organisation for Applied Scientific Research (TNO)	Lead: WCH	Lead: WCH Support: AIR/TNO
	Uganda process evaluation	Lead: AIR Support: WCH	Lead: AIR Support: WCH	Lead: AIR
World University Service of Canada (WUSC), Equity in Education in Refugee Camps in Kenya (EERCK) programme	Impact evaluation in Kakuma	Lead: AIR Support: WUSC	Lead: WUSC Enumerator training: AIR	Lead: AIR Inputs: WUSC
	Impact evaluation in Dadaab	Lead: AIR Support: WUSC	Lead: WUSC Enumerator training: AIR	Lead: AIR Inputs: WUSC
	Process evaluation	Lead: AIR Support: WUSC	Lead: WUSC Enumerator training: AIR	Lead: AIR Inputs: WUSC
Kepler (Kepler University programme)	Process evaluation	Lead: AIR Support: Kepler	Lead: AIR Support: Kepler	Lead: AIR Inputs: Kepler
	Employer survey	Lead: AIR Support: Kepler	Lead: Local data collection firm, Laterite Enumerator training: AIR	Lead: AIR Inputs: Kepler
	Costing analysis	Lead: AIR Support: Kepler	Lead: AIR	Lead: AIR
Libraries Without Borders (LWB),	Impact evaluation of self-managed Ideas Box	Lead: AIR Support: LWB	Lead: LWB	Lead: AIR Support: LWB
	Impact evaluation of Ideas Box in DRC centre	Lead: AIR Support: LWB	Lead: DRC Inputs: AIR	Lead: AIR Support: LWB

Innovation Team (Programme)	Evaluation	Evaluation Design	Data Collection	Data Analysis
Ideas Box programme	Process evaluation	Lead: AIR Support: LWB	Lead: AIR Support: LWB	Lead: AIR Support: LWB
Caritas, Essence of Learning (EoL) programme	Process evaluation in Romania	Lead: AIR Support: Caritas	Lead: AIR Support: Caritas	Lead: AIR Support: Caritas
	Process evaluation in Bangladesh	Lead: AIR Support: Caritas	Lead: AIR Support: Caritas	Lead: AIR Support: Caritas

Annex B. Kepler Cost Analysis and Employer Survey

Exhibit B-1. Kepler Cost Analysis

ACTIVITY	Cost Category	Cost, Kiziba ($)	Cost, Replacement ($)	Savings ($)
Negotiate and sign a lease on staff/meeting house	Personnel	1083.33	176.92	906.41
Manage the process of setting up an electricity supply and Internet access for the Kiziba campus	Contracted services	52956.75	42561.23	10395.52
Create foundational year courses (including 21st Century Communications, Technology Skills, Professional Competencies, and Book Clubs/Newsela Lessons) to equip Kiziba students to excel in Kepler courses	Personnel	21538.46	3076.92	18461.54
Create a 6-week introductory course, in the absence of the Internet, to orient students to academic requirements	Personnel	32000	3076.92	28923.08
Develop a curriculum tailored to Kiziba's needs that combines offline and online elements	Personnel	6576.92	3000	3576.92
Train Kepler Kiziba curriculum associate to modify and create Kiziba-specific materials	Personnel	4923.076	369.23	3923.84
			Total Savings	**66187.31**

Exhibit B-2. Employer Survey—Likelihood of Hiring

	Likelihood of Hiring	Likelihood of Hiring
Refugee	-6.621***	-8.512***
	(1.340)	(2.372)
Experience		6.155***
		(0.438)
Female		-12.18***
		(1.887)
Refugee x Experience		-4.163***
		(0.776)
Refugee x Female		12.88***
		(3.138)
Constant	64.33***	61.21***
	(1.221)	(1.445)
Observations	1702	1702
R^2	0.014	0.101
Adjusted R^2	0.012	0.099

Standard errors in parentheses
* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Annex C. Impact of War Child Holland's Can't Wait to Learn Programme in Jordan and Sudan

Exhibit C-1. Baseline Summary Statistics and Balance in Jordan

Variables	N	Control Mean	Treatment Mean	SMD
Panel A: Child Characteristics				
Age	707	7.15	7.54	-0.34†
		(1.13)	(0.32)	
Female child	707	0.63	0.55	-0.16
		(0.48)	(0.11)	
Number of siblings	707	4.20	4.23	-0.02
		(1.64)	(0.21)	
Helps with childcare	707	0.41	0.41	0.01
		(0.81)	(0.04)	
Jordanian	707	0.81	0.78	0.07
		(0.40)	(0.09)	
Syrian	707	0.15	0.19	-0.10
		(0.36)	(0.09)	
Lives in refugee camp	707	0.09	0.11	-0.08
		(0.28)	(0.08)	
Children in household earn income	707	0.29	0.71	-0.28†
		(1.22)	(0.23)	
Panel B: Education Background				
Spent at least 1 month out of school	707	0.02	0.02	-0.14
		(0.13)	(0.01)	
School grade	707	1.90	2.24	-0.41†
		(0.84)	(0.27)	
Panel C: Parent Characteristics				
Mother completed primary school	707	0.67	0.67	-0.00
		(0.47)	(0.08)	
Father completed primary school	707	0.68	0.61	0.14
		(0.47)	(0.07)	

Variables	N	Control Mean	Treatment Mean	SMD
Mother completed secondary school	707	0.43	0.40	0.05
		(0.50)	(0.07)	
Father completed secondary school	707	0.38	0.38	0.01
		(0.49)	(0.06)	
Mother engaged in income-generating activity	707	0.13	0.13	-0.00
		(0.34)	(0.04)	
Father engaged in income-generating activity	707	0.70	0.66	0.08
		(0.46)	(0.05)	
Household income (binned)	698	2.24	2.19	0.08
		(0.72)	(0.09)	

Notes: Standard deviation and standard error reported in brackets for the third and fourth columns, respectively. Balance coefficients report the regression coefficient for children in the treatment group relative to the comparison group. Significance at the 90%, 95%, and 99% confidence intervals are indicated by *, **, and ***, respectively. † indicates standardised mean difference greater than 0.2. All regressions control for school-type strata (single shift, owned; single shift, rented; double shift, owned; double shift, rented). Variables for grade, age, and gender come from assessment surveys. Monthly income reported as a categorical variable: 1=less than 100 JD, 2=101–400 JD, 3=401–700 JD, 4=701–1,000 JD, and 5=more than 1,000 JD.

Exhibit C-2. Impact Estimates: Academic Outcomes in Jordan

	N	Pre-Programme Mean	Treatment Coefficient
Arabic	1377	-0.000	0.053
			(0.101)
Mathematics	1377	0.000	-0.024
			(0.070)

Note: Clustered standard errors reported in parentheses. Significance at 90%, 95%, and 99% confidence levels indicated by *, **, and ***, respectively. Ordinary least squares (OLS) regressions reported. All regressions control for school-type strata (single shift, owned; single shift, rented; double shift, owned; double shift, rented) and unbalanced covariates (age, grade, and whether children contribute financially to the household).

Exhibit C-3. Impacts on Attendance in Jordan

	Attendance (Days Per Week)		
	[1]	[2]	[3]
Treatment	-0.096	-0.085	-0.11
	(0.078)	(0.076)	(0.104)
Constant	4.005***	4.028***	3.806***
	(0.039)	(0.063)	(0.267)
Observations	696	696	694
R^2	0.008	0.016	0.031
Additional Controls			
Strata		✓	✓
Unbalanced covariates			✓

Note: The dependent variable is an indicator variable equal to one if the study participant attrited during the study. Strata include single shift, owned; single shift, rented; double shift, owned; and double shift, rented. Unbalanced covariates include age, grade, and whether children contribute financially to the household.

Exhibit C-4. Impact Estimates: Psychosocial Outcomes in Jordan

	N	Pre-Programme Mean	Treatment Coefficient	Westfall-Young Adjusted p-Value
Hope	1360	0.000	0.229**	0.044
			(0.103)	
Rosenberg self-esteem	1366	0.000	-0.018	0.846
			(0.078)	
Child self-efficacy	1364	-0.000	0.111	0.429
			(0.091)	
Stirling well-being	1377	-0.000	-0.053	0.731
			(0.095)	

Note: Clustered standard errors reported in parentheses. Significance at 90%, 95%, and 99% confidence levels indicated by *, **, and ***, respectively. OLS regressions reported. All regressions control for school-type strata (single shift, owned; single shift, rented; double shift, owned; double shift, rented) and unbalanced covariates (age, grade, and whether children contribute financially to the household).

Exhibit C-5. Impact Estimates on Arabic and Mathematics Sub-Component Scores in Jordan

	N	Pre-Programme Mean	Treatment Coefficient
Panel A: Arabic Sub-Components			
Concept of print	1377	0.000	-0.190*
			(0.103)
Phonological awareness	1377	0.000	0.042
			(0.111)
Vocabulary	1377	0.000	0.129
			(0.091)
Listening comprehension	1377	0.000	-0.065
			(0.086)
Letter naming	1377	0.000	0.144
			(0.124)
Letter sounds	1377	0.000	-0.072
			(0.126)
Reading fluency	1377	0.000	0.019
			(0.085)
Reading comprehension	1377	-0.000	0.118
			(0.135)
Writing	1377	-0.000	0.027
			(0.090)
Panel B: Mathematics Sub-Components			
Addition	1362	-0.000	-0.021
			(0.088)
Subtraction	1362	-0.000	-0.026
			(0.100)
Division	1362	0.000	0.140
			(0.101)
Multiplication	1362	-0.000	-0.120
			(0.109)
Number recognition	1362	-0.000	-0.019
			(0.070)

	N	Pre-Programme Mean	Treatment Coefficient
Quantity recognition	1362	0.000	0.036
			(0.077)
Time	1362	-0.000	-0.127
			(0.103)
Shapes	1362	-0.000	0.082
			(0.100)

Note: Clustered standard errors reported in parentheses. Significance at 90%, 95%, and 99% confidence levels indicated by *, **, and ***, respectively. OLS regressions reported. All regressions control for school-type strata (single shift, owned; single shift, rented; double shift, owned; double shift, rented) and unbalanced covariates (age, grade, and whether children contribute financially to the household).

Exhibit C-6. Impact on Literacy Outcomes in Sudan

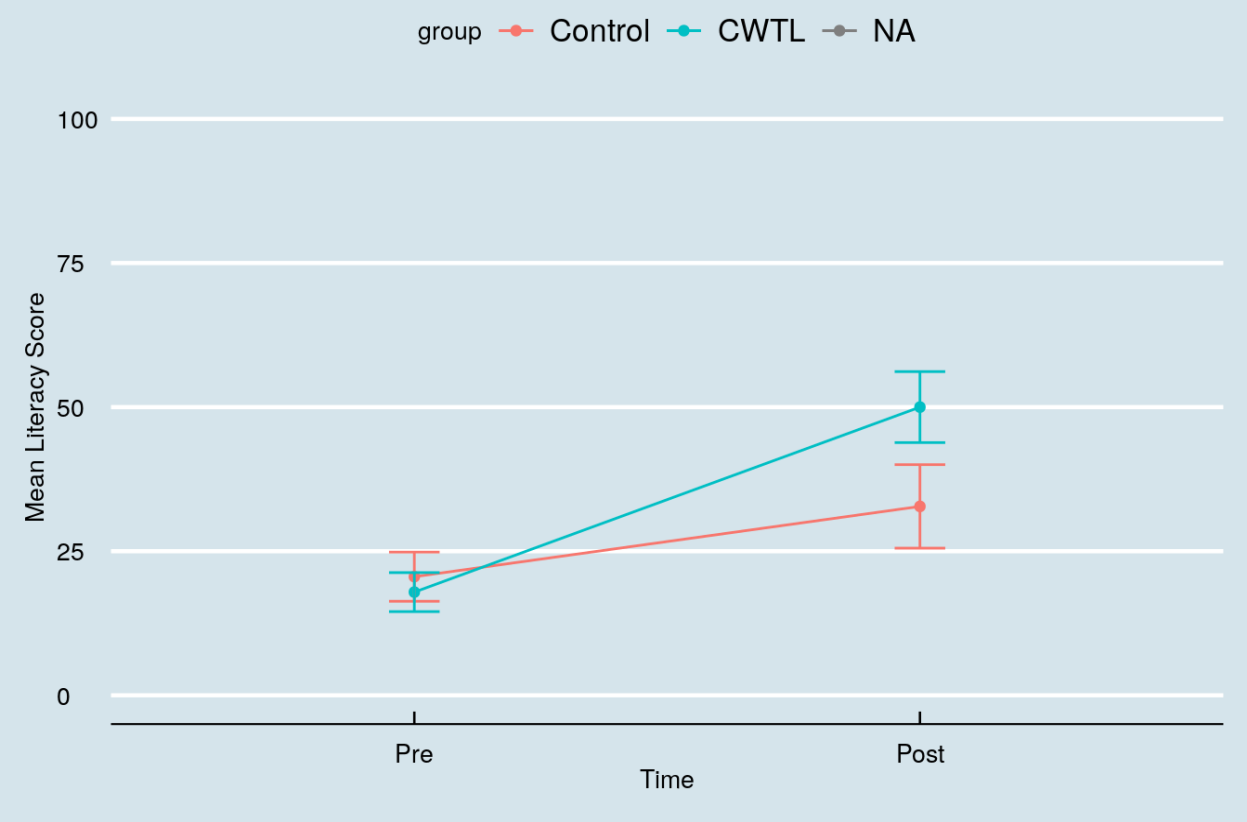
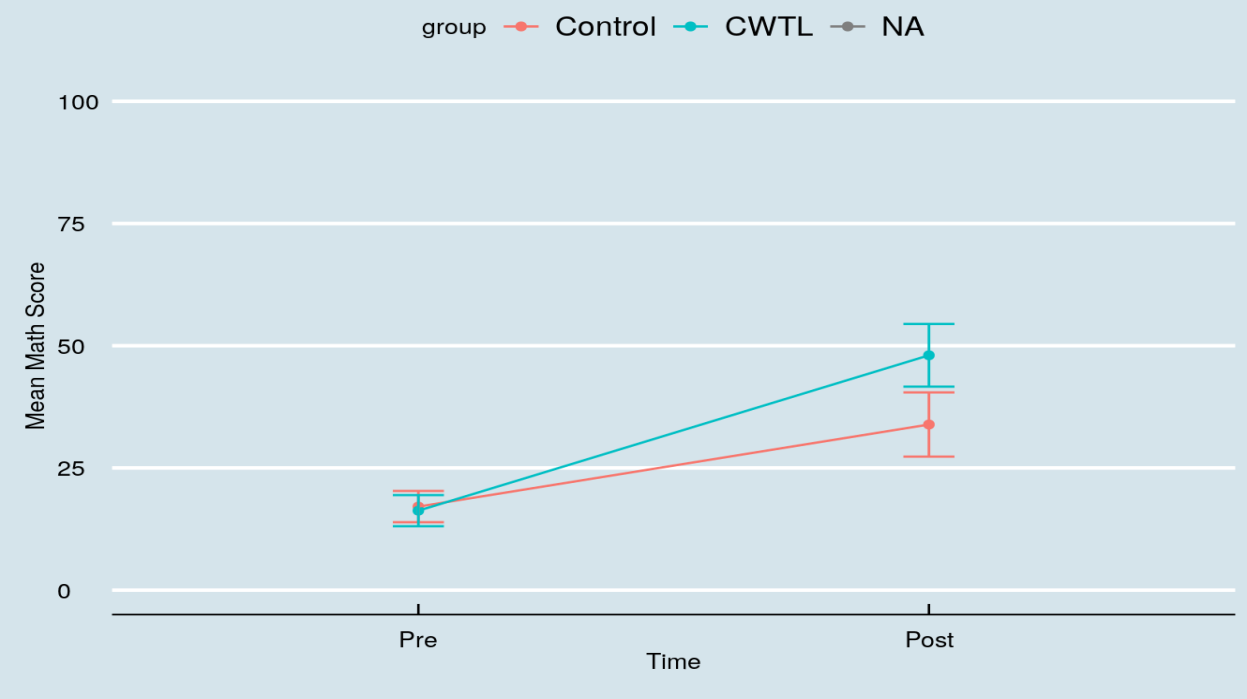


Exhibit C-7. Impact on Mathematics Outcomes in Sudan



Annex D. Impact of the World University Service of Canada's Equity in Education in Refugee Camps in Kenya Programme

Exhibit D-1. Impact Estimates: Academic Outcomes in Kakuma

	N	Treatment Effect
Standardised Kenya Certificate for Primary Education (KCPE) results	435	-0.917
		(5.215)
Uwezo results	1152	-0.483
		(0.262)

Note. Robust standard errors in parentheses. * $p < 0.05$. ** $p < 0.01$. *** $p < 0.001$.

Exhibit D-2. Impact Estimates: Attendance in Kakuma

	N	Treatment Effect
Primary school days attended	928	0.636
		(2.928)

Note. Robust standard errors in parentheses. * $p < 0.05$. ** $p < 0.01$. *** $p < 0.001$.

Exhibit D-3. Impact Estimates: Psychosocial Outcomes in Kakuma

	N	Treatment Coefficient
Aspirations index	1106	-0.118
		(0.0649)
Resilience index	1106	-0.0861
		(0.0497)

Note. Robust standard errors in parentheses. * $p < 0.05$. ** $p < 0.01$. *** $p < 0.001$.

Exhibit D-4. Impact Estimates: Academic Outcomes in Dadaab

	N	Treatment Coefficient
Standardised KCPE results	161	-10.57
		(21.41)
Uwezo results	532	0.032
		(0.113)

Note. Robust standard errors in parentheses. * $p < 0.05$. ** $p < 0.01$. *** $p < 0.001$.

Exhibit D-5. Impact Estimates: Attendance in Dadaab

	N	Treatment Coefficient
Primary school days attended	585	1.278
		(3.064)

Note. Robust standard errors in parentheses. * $p < 0.05$. ** $p < 0.01$. *** $p < 0.001$.

Exhibit D-6. Impact Estimates: Psychosocial Outcomes in Dadaab

	N	Treatment Coefficient
Aspirations index	213	0.0107
		(0.440)
Resilience index	213	0.192
		(0.374)

Note. Robust standard errors in parentheses. * $p < 0.05$. ** $p < 0.01$. *** $p < 0.001$.

Annex E. Impact of Libraries Without Borders' Ideas Box Programme

Exhibit E-1. Impact of the Ideas Box Using Difference-in-Difference Analysis and Propensity Score Matching

	(1) Overall Average	(2) Prosocial Behaviour	(3) Cognitive and Emotional Functioning	(4) Daily Tasks and Problem-Solving	(5) Self-Esteem	(6) Child Protection
Psychosocial outcomes	-0.213***	-0.103*	-0.324***	-0.240***	-0.200***	-0.199***
	(0.0378)	(0.0477)	(0.0558)	(0.0527)	(0.0456)	(0.0578)
Observations	1084	1084	1084	1084	1084	1084

Annex F. Common Coding Structure

I. **Context:** This is the parent node of the following:

- Security
- Gender norms
- Access and physical resources
- Technology
- Legal rights
- Cultural norms
- Social norms

Context is one of the three conceptual domains of the scaling research for the meta-evaluation. In exploring findings related to context, we hope to identify which contextual factors (security situation, gender norms, etc.) enable or inhibit programme implementation and/or the effects of the five programmes. What lessons learned about these contextual factors can be applied in other contexts or when the programmes are scaled up? What key questions need to be asked to inform scale-up and other key decisions?

a. Security: Security is a contextual factor that refers to the security situation of the specific geographic location (city/district/province/country) or context in which the programme is being implemented (or potentially implemented), the security or future security of programme implementers and beneficiaries, the security or future security of programme infrastructure/materials, etc.

b. Gender norms: Gender norms are a contextual factor that refers to the norms around gender that exist in different implementation contexts and how these genders norms may influence programme implementation and impact (e.g., do gender norms moderate or mediate the implementation, scaling, and/or impact of the programme?). For example, how do the gender norms in Sudan influence how the Can't Wait to Learn (CWTL) programme is designed and implemented there? How do gender norms in Jordan influence how the Ideas Box programme is designed and implemented there?

c. Access and physical resources: Access and physical resources are contextual factors that refer to the amount and type of access that implementers and beneficiaries have to current or future programme implementation sites, as well as the access implementers have to the physical resources (including infrastructure and materials) needed to implement their programme. This includes not only the implementing agency itself but also current and future partners. For example, War Child Holland (WCH) staff have limited access to CWTL implementation sites in Sudan because they are not typically cleared to travel outside Khartoum. A related example from CWTL is the programme's ability to procure tablets in each of the countries in which it works.

d. Technology: Limited or restricted programme implementation in protracted humanitarian crisis settings can be bolstered through the use of various forms of information and

communication technology. For example, technology can be used as a direct educational tool (i.e., providing students with cell phones, tablets, etc. to use as a learning tool) or to bolster programmatic capacity (e.g., remote teacher training, etc.). Data that demonstrates the use of different kinds of integrated communication technology (ICT; or the barriers to and facilitators of their use) should be included in this node.

e. Legal rights and institutional structures: In many countries, refugees have restrictive legal, social, and economic rights that can limit their access to education and other social services and prohibit or limit their ability to engage in the labour market. Limited legal rights may be the result of national, state, or local laws that restrict non-citizens' (or refugee/migrant populations') access to education, employment, or other social services that are critical to programme implementation.

f. Cultural norms: Fehr and Gächter (2000) define a social norm as "(1) a behavioural regularity that is (2) based on a socially shared belief of how one ought to behave, which triggers (3) the enforcement of the prescribed behaviour by informal social sanctions."

g. Social exclusion: We will capture systematic social exclusion of groups, which is not necessarily captured in "norms." For example, the Roma population in Romania faces prejudice that shapes their access to services, but this prejudice is not necessarily a social "norm."

h. Future prospects for participants: What do programme implementers hope to achieve in practical terms for clients/participants—for example, labour market participation, secondary school enrolment? Is there any information that could help to answer questions about whether the programme is participant-focused—for example, is the programme appropriate given the labour market? Are secondary schools available, so that enrolment can be a primary aim of a primary school programme?

II. **Ownership and Advocacy:** This is a parent node. Ownership and advocacy is one of the three conceptual domains of the scaling research for the meta-evaluation. According to the literature on education in refugee settings, local buy-in and ownership of programmes can greatly facilitate scale-up in refugee contexts. For this reason, enhancing community support and ownership of a programme should be a key objective. The three nodes that fall under the theme of "ownership and advocacy" are local demand, political buy-in, and community support.

a. Demand: Demand refers to any information about how the implementers identified the demand or need for the programme, or how implementers created demand if it did not exist. Research has shown that programmes that stimulate demand for education are most effective in improving school enrolment and attendance (de Hoop et al., 2018). Conversely, an increase in demand could create challenges for implementation if there are not enough resources to meet the level of demand. Any challenges associated with demand will also be included in this node.

b. Political buy-in: Political buy-in refers to any information about how programmes negotiate to gain permission to work in a certain physical or conceptual space. Political sensitivities could affect whether scale-up is possible or desirable. Political will amongst host governments (at local and national levels, and amongst other unofficial governance

structures in a refugee space) must exist for programmes to operate effectively. Political economy concerns associated with integrating programmes into national education systems can reduce the effectiveness of education programmes at scale. Programme flexibility and close engagement with the government can, in many cases, mitigate issues associated with integrating programmes into national education systems (de Hoop et al., 2018).

c. Community support: Similar to political buy-in, community support refers to any information about how programmes negotiate permission amongst influential community members or parties (e.g., parents, in the case of education) to implement a programme, or any challenges within the community that inhibited or might inhibit programme implementation, or any strategies to involve community members in programming to create buy-in and ownership. Rejection of the presence of a programme or of certain elements of programme implementation could act as a barrier to implementation (de Hoop et al., 2018). In addition, a programme that does not seek community support before implementation risks community backlash against programme beneficiaries, not just implementers (Chinen, Coombes, de Hoop, Elmeski, & Castro Zazur, 2017).

III. **Business Model:** We take this domain from Ramalingam et al. (2015) and Obrecht et al. (2017), who discuss the need for humanitarian innovations to incorporate business plans and processes that will guide implementation. Assessing the extent to which teams have incorporated a business model into project design and thinking will include (but is not limited to) elements such as: (a) financial resources (Ramalingam et al., 2015); (b) project management; (c) project personnel; (d) procurement; and (e) partnerships. Information included under this domain should answer the following questions: How did the programme arrive at its current business model (including partnership models, expansion processes, finances, staffing, management, etc.)? What criteria were used to select the business model? How is the business model re-evaluated throughout the scaling process? How flexible is the business model to changes in programme implementation after scale-up?

a. Financial resources: This refers to information about how programmes have allocated financial resources during the innovation stage of their programming and adapted lessons learned for financing implementation. It also includes any information about funding uncertainty for a programme moving forward. Including cost-effectiveness estimates in impact evaluations may help to encourage governments and donor agencies to support the scale-up of education programmes (de Hoop et al., 2018). Any information on lessons learned about financing and procuring funds or in-kind contributions for running programmes (e.g., if the physical space or other services were donated) should be included in this node.

b. Organisational management: This refers to the way in which the innovation teams address resource, information, and capacity gaps, and how they make decisions to adapt their models over time (Pritchett, Samji, & Hammer, 2012). In this node, we are interested in understanding how senior and mid-level staff on the innovation teams bring focus and leadership, how they decide to grow the organisation or manage the change process, and how they learn from failure.

c. Project personnel: This refers to the capacities of field and operational staff, and how innovators build staff and team capabilities and "identify and train the relevant skills in individuals and groups to enable innovation to take place" (Bessant et al., 2014, p. 32). Under this node, we can code how respondents discuss professional development (training, coaching, mentorship, resources provided to the implementers of the programmes).

d. Procurement: This refers to the goods or services that are used by the innovation teams, as well as the processes that innovators use to obtain them. Betts and Bloom (2014, p. 10) state that there is an assumption that "humanitarian goods can only come from a closed and tightly regulated group of suppliers. Inter-agency coordination and procurement tend to privilege a small group of mainly UN organisations and international NGOs, whether or not they are the most efficient or effective of providers. These organisations may, in turn, privilege known suppliers rather than reaching out to alternative solution holders." Under this node, we can code how innovators procure goods and services to explore if they use alternative models and follow procurement processes based on "performance and value, opening the system to non-traditional ideas and suppliers" (Betts & Bloom, 2014, p. 11).

e. Partnerships: This refers to the ways in which innovation teams develop and maintain relationships "within aid agencies, between international and national actors, with academics, with the private sector, and between operational agencies" (Bessant et al., 2014, p. 27).

f. Exit strategy: This includes any information about how the implementer plans to hand over the programme to local implementers and phase out their own involvement. Given the lack of reliable and consistent funding for project implementation, programmes have expressed the need to incorporate activities with the government or with other structures within refugee camps into country programming.

General Sub-Nodes

a. Barriers: These are the characteristics under each component of the three scaling domains (scaling as a knowledge question, ownership and advocacy, and business model) that impede or reduce the ability of the innovation team to scale up their education programme in a protracted humanitarian setting.

b. Facilitators: These are the factors that help to overcome challenges that may arise during the scaling process by easing or limiting the role of the noted barriers (security, gender norms, cultural norms, access and physical resources, technology, and legal rights) and/or are necessary for delivering—or can increase—the positive impacts of innovations in education. For example, signing memoranda of understanding (MOUs) with partners enables Libraries Without Borders (LWB) to have stronger partnerships, which could be seen as a facilitator of partnerships. Similarly, the presence of high-quality teachers could increase the effects of innovations in education.

c. Adaptation: This refers to responses to changes in circumstances and to new information about the efficacy of the programme (Valters, Cummings, & Nixon, 2016). For example, while LWB's Ideas Box typically relies on electricity and Internet access, LWB has made modifications so that the box still works in places that lack constant access to electricity

(by installing a generator in the Ideas Box) and the Internet (by enabling the Ideas Box to work offline).

d. Description: This refers to any descriptive information related to the parent node that is not captured by the other sub-nodes but is still important for the purposes of the process evaluation.

e. Recommendations: These are suggestions for how the innovation teams could be more responsive to barriers to scale-up, stimulate facilitators towards scaling up, or make improvements related to the innovation or organisation itself. For example, the idea that LWB should pursue a partnership with Save the Children at the global level (instead of developing partnerships on a country-by-country basis) could be coded as a recommendation under partnerships.

Annex G. Ethical Approvals

AIR complies with best practices in the area of ethical research. It is a registered institution with the Office for Human Research Protections and has signed an assurance statement confirming that we will abide by federal regulations.

All projects that have data collection plans involving human subjects are thoroughly reviewed at the programme level and at the institutional level before data collection can proceed in the field. AIR's Institutional Review Board (IRB) also ensures that all data security measures are in place and have been cleared through AIR's security director, and that AIR meets ethical and legal standards in its work. The IRB includes a certified IRB chair, who spells out the responsibilities of the researcher to protect the rights and privacy of evaluation participants.

AIR's IRB follows the standards set forth by the American Evaluation Association Guidelines and the Joint Committee on Standards for Educational Evaluation. These standards can be distilled into three general principles: (1) evaluators will conduct evaluations legally and ethically, taking into account the welfare of those involved in the evaluation, as well as the general public; (2) evaluators will conduct evaluations in a competent and efficient fashion that will lead to reliable and accurate results; and (3) evaluators will design evaluations and report the results in a manner that is useful and appropriate for the intended audience. There are also clear guidelines regarding the expectations with which local data collectors must comply (e.g., how to document informed consent, how to store and restrict access to physical files and electronic data files, and how to handle identifiable information).

AIR usually requests IRB approval from a local institution, but complex emergency settings may not always have IRB systems in place, or there may not be sufficient time before baseline data collection to secure local IRB approval, preventing AIR from submitting a request to a local institution. We have obtained local IRB approval in Lebanon, Kenya, Rwanda, and Sudan. In cases where applying for local IRB approval is not feasible, AIR follows the Code of Conduct of the United Nations Evaluation Group (UNEG), which requires both a conflict- and gender-sensitive approach to research, adherence to the "do no harm" principle, as well as transparency, confidentiality, accuracy, accountability, and reliability, amongst other key principles (United Nations Evaluation Group, 2008). Specifically with regard to the protection of vulnerable individuals and communities, AIR respects and adheres to the United Nations (UN) Declaration of Human Rights, the UN Refugee Convention, the Convention on the Rights of the Child, and the Convention on the Elimination of all forms of Discrimination Against Women, as well as other human rights conventions and national legal codes that respect local customs and cultural traditions, religious beliefs and practices, personal interaction, gender roles, disability, age, and ethnicity (United Nations Evaluation Group, 2008).

Annex H. Methodology for the Initial and Updated Evidence Synthesis

Initial Evidence Synthesis

To identify the relevant literature on effectively scaling up education programmes in refugee settings, we relied on a rapid but systematic approach. We began by summarising the relevant literature identified by the United Nations Children's Fund (UNICEF) before the start of our literature review. Specifically, UNICEF provided AIR with 19 studies that it considered relevant to scaling up education programmes in refugee settings. Simultaneously, we developed a search strategy to identify other relevant articles from peer-reviewed journals and grey literature, including implementation reports from humanitarian organisations. The latter included project documents describing the background, design, and implementation features of a programme; process evaluations assessing whether a programme was being implemented as intended and how it was working; and impact evaluations designed to determine the effects of a programme (using either experimental or quasi-experimental methods). Finally, we identified examples of scaled-up programmes discussed in previous overviews of the literature. In addition, we included literature identified by the Reference Group for the HEA following a review of a first draft of the scaling report.

We did not seek to perform an exhaustive review of the literature. Instead, we relied on a combination of formal systematic search approaches and our knowledge of the relevant literature to synthesise the pertinent evidence. This combination of formal, systematic, and informal approaches enabled us to provide an overview of the main barriers to and facilitators of effectively scaling up education programmes in refugee settings. It is possible that relevant evidence has not been included due to the rapid nature of our review. To mitigate this concern, updated our review over the course of the Humanitarian Education Accelerator (HEA). As part of the updates, we also incorporated the impact and process evaluations AIR conducted under the HEA. In the final update of this review, we synthesised the evidence on how to scale up education programmes in refugee settings through a narrative synthesis.

The searches we conducted aimed to identify several types of evidence that could inform the scale-up of education programmes in refugee settings, including research and information about scale-up processes from implementing organisations and other key stakeholders. In order to be included in our review, research and implementation evidence needed to focus on one of the following: (1) education programmes that were successfully scaled up in refugee settings, (2) education programmes that were successfully scaled up outside refugee settings, or (3) non-education programmes that were successfully scaled up in refugee settings. We also included studies on pilot education programmes in refugee settings that provided impact evaluation evidence of programme effectiveness but no evidence that these programmes had been successfully scaled up. This impact evaluation evidence could come from both experimental and non-experimental studies.

Although we did not conduct a systematic review, we relied on established best practices for conducting systematic reviews. These practices allowed us to identify a wide range of studies and implementation documents that are relevant to scaling up education programmes in refugee settings. Specifically, we pre-specified the databases in which we would search for evidence, along with the search terms we would use

(selected based on our understanding of the relevant literature). We also conducted citation tracking of key papers identified by UNICEF. Finally, we included studies and implementation documents that we considered relevant, based on our knowledge of impact and process evaluations, as well as our knowledge about implementing education programmes in refugee settings. In the sections that follow, we describe the procedures for our review in more detail.

Search

In addition to reviewing the literature provided by UNICEF, we identified additional peer-reviewed literature by searching several databases, including EBSCO, JSTOR, ProQuest, Web of Science, and CIAO. We also conducted Internet searches using Google Scholar. The search strings were designed to return studies that included thematic, population, technical, programmatic, and geographic terms relevant to the scale-up of education programmes in refugee settings. We did not rely solely on literature related to education or refugee settings; we also included evidence on the scale-up of education programmes outside refugee settings and the scale-up of non-education programmes in refugee settings. We took these choices into consideration when defining our keywords, which are summarised below:

- Thematic terms: “education/livelihoods/ICT/health”
- Population terms: “refugee/urban refugees/asylum seekers/IDPs/displaced”
- Technical terms: “scale/scale up/scale out/case study,” “pilot/evaluation/study/experiment”
- Programmatic terms: “project/programme/intervention”
- Geographic terms: “low-income countries,” “middle-income countries,” “developing,” “underdeveloped,” “less developed”

Forward and backward snowballing of the references cited in papers identified by UNICEF provided additional peer-reviewed studies that may not have been found in database searches.

Updated Evidence Synthesis

For the updated evidence synthesis, we conducted searches that aimed to identify several types of evidence published after 2017 that could inform the scale-up of education programmes in refugee settings, including research and information about scale-up processes from implementing organisations and other key stakeholders. To identify relevant evidence, we used the following process.

Step 1

Identify peer-reviewed literature by searching several databases, including EBSCO, JSTOR, ProQuest, Web of Science, and CIAO. Conduct Internet searches using Google Scholar and the following search strings:

- Thematic terms: “education/teacher/remedial/ICT/technology/teacher training”
- Population terms: “refugee/urban refugees/asylum seekers/IDPs/displaced/humanitarian/ fragile states”
- Technical terms: “scale/scale up/scale out/case study,” “pilot/evaluation/study/experiment”
- Programmatic terms: “project/programme/intervention”

- Geographic terms: “low-income countries,” “middle-income countries,” “developing,” “underdeveloped,” “less developed”

Construct search terms using one term from each category—for example, “education + refugee + scale + project + low-income countries.”

Classify the results as follows:

- Results from the thematic term “remedial” go into the folder “Remedial.”
- Results from the thematic term “teacher” go into the folder “Teacher Training.”
- Results from the thematic terms “ICT” and “technology” go into the folder “Technology.”
- Results on how education programmes move to scale go into the folder “Scale.”

Files should be named with the first five words of the title.

Stop at 10 pages of results in each database.

Step 2

Screen in two phases: first on the basis of titles and abstracts, and then on the basis of full texts. For each title/abstract, in the Excel file tabulating the results attach the following labels as appropriate: “research,” “implementation,” “impact study,” “positive impact,” “refugee context,” “education,” “scaled,” and “other.”

Step 3

After attaching labels to each article, conduct full-text reviews and write summaries of each included article. Focus on the methods section (to ensure it accurately describes what the paper claims to be) and the programme description section (to ensure that it is indeed about remedial education/teacher training/scaling in the appropriate setting).

Step 4

In addition to the research evidence, conduct a targeted search (using the same or similar key words) of the websites of several humanitarian organisations and research organisations (see Exhibit H-1) with a focus on education in humanitarian settings.

Exhibit H-1. Organisations With a Focus on Education in Humanitarian Settings

Organisation
Education Development Trust
Overseas Development Institute
Brookings
World Bank

Organisation
Refugees Studies Centre
International Organisation for Migration (IOM)
Norwegian Refugee Council
Save the Children (UK)
Queen Rania Organisations
United Nations Relief and Works Agency (UNRWA)
The Roma Education Fund
Organisation for Eelam Refugee Rehabilitation (OfFER)
Global Innovation Fund
Development Innovation Ventures
CARE
International Federation of Red Cross and Red Crescent Societies
Mercy Corps
Danish Refugee Council
International Rescue Committee (IRC)
United Nations Refugee Agency (UNHCR)
United Nations Children's Fund (UNICEF)

Annex I. Summary of Methods Used for the Individual Impact Evaluations

Impact Evaluation of the World University Service of Canada's Equity in Education in Refugee Camps in Kenya Programme

AIR designed a two-part study to evaluate the impact of the Equity in Education in Refugee Camps in Kenya (EERCK) programme on cognitive and non-cognitive skills for seventh- and eighth-grade girls in the 2017 and 2018 cohorts. In Kakuma, we used a randomisation-by-oversubscription methodology, in which we randomly assigned 1,293 programme-eligible girls to receive the programme and 983 to a control group that did not receive the programme. In Dadaab, we used a regression discontinuity design to determine the impact of the programme. We compared the outcomes for 988 girls who were above the cut-off for admittance into the programme with the outcomes for 418 comparison girls who were just below the cut-off for admittance. Exhibit I-1 details the sample size for each research design.

Exhibit I-1. Differences in Evaluation Design

Kakuma	Dadaab
Sample size: 2,457 girls distributed across the treatment and control groups	Sample size: 988 treatment girls and 418 comparison girls
Surveyed sample: 1,293 treatment girls and 983 control girls	Surveyed sample: 825 treatment girls and 358 comparison girls

Impact Evaluation of War Child Holland's Can't Wait to Learn Programme in Jordan and Sudan

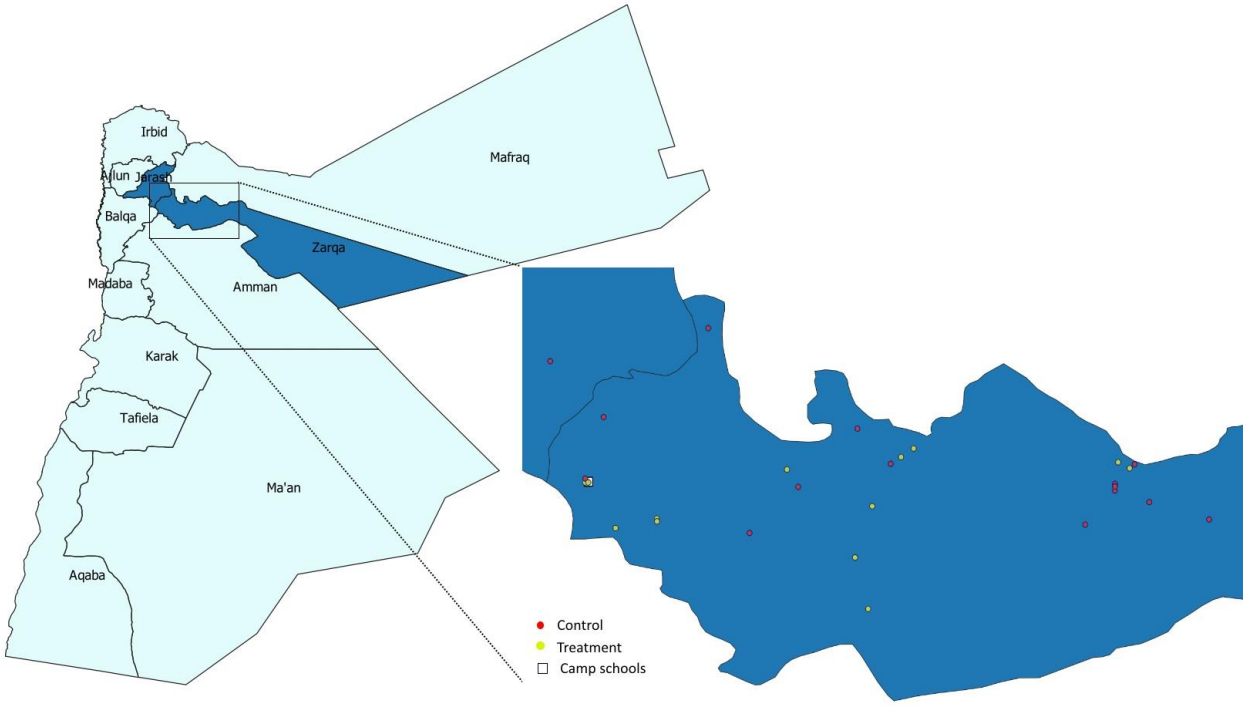
We measured the impact of the Can't Wait to Learn (CWTL) programme on reading and numeracy outcomes (reading outcomes only in Jordan), as well as psychological well-being. War Child Holland (WCH) developed the mathematics and literacy assessments based on a didactic framework that was developed in line with the learning objectives of the national curricula in Jordan and Sudan, existing early grade reading assessments (EGRA), and existing early grade mathematics assessments (EGMA). The mathematics assessment examines children's knowledge of number and place value, addition and subtraction, multiplication and division, time, and shapes. The reading assessment examines children's knowledge of print, oral literacy skills, decoding skills, vocabulary, and writing. To measure children's psychological well-being, we used a battery of tests including the Stirling Children's Well-Being Scale, the Children's Hope Scale, an adaptation of the Rosenberg Self-Esteem Scale, and measures of self-efficacy, motivation, and future orientation developed by the Netherlands Organisation for Applied Scientific Research (TNO) for use in the earlier research studies in Sudan.

Study Design in Jordan

We used a quasi-experimental difference-in-difference analysis with 17 treatment and 18 comparison schools to determine the effectiveness of the programme compared to traditional education in Jordan. This approach produces valid impact estimates under the assumption that treatment and comparison groups show parallel trends in the outcome variables in the absence of the programme. Although we cannot test this assumption, the likelihood of parallel trends increases when treatment and comparison groups are selected based on transparent and observable selection criteria, which minimises the likelihood of selection bias.

To achieve this goal, we assigned schools to the treatment and comparison groups based on geographic and school characteristics. This approach helped us to select treatment and comparison schools that were similar in observable geographic and school characteristics. Exhibit I-2 shows the geographical distribution of treatment and comparison schools.

Exhibit I-2. Distribution of Treatment and Comparison Schools



From each school, we selected a maximum of 30 students in one class (class section A, selected from classes in Grades 1–3) to take part in the research. In classrooms with more than 30 students, we randomly selected 30 students to take part in the evaluation.

Study Design in Sudan

We used a quasi-experimental difference-in-difference analysis with four treatment villages and four comparison villages to determine the effectiveness of the programme compared to a traditional education centre for out-of-school children in Sudan. Treatment and comparison villages were eligible to participate

in the programme if the community agreed to participate, each village had at least 20–30 children who had never been to school, the travel distance to the nearby cities of Kassala and Sinnar was not prohibitive for facilitating programme implementation, the village was safe to access during day and night, and the community was willing to receive visitors and ensure the safety and protection of programme materials. Treatment villages were only eligible if no school or educational centre could be found within a radius of 3 kilometres of the village, and comparison villages were only eligible if they included a traditional education centre for out-of-school children. The difference-in-difference approach produces valid impact estimates under the assumption that treatment and comparison groups show parallel trends in the outcome variables in the absence of the programme. Data collected before the baseline survey demonstrated parallel trends across treatment and comparison group assessments before the start of the programme.

Impact Evaluation of Libraries Without Borders' Ideas Box Programme

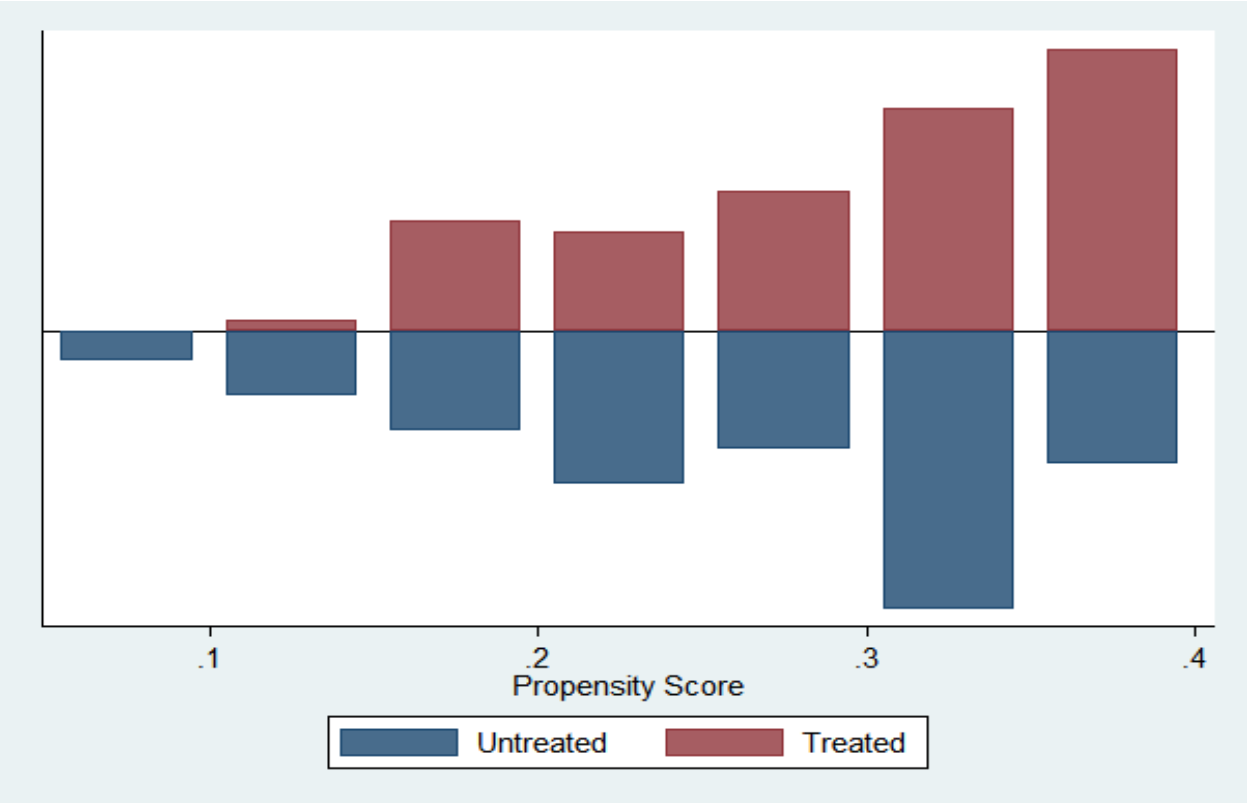
We conducted one non-experimental impact evaluation and one proof-of-concept study to assess the effects of the Ideas Box programme: (1) an evaluation to determine the additive effects of Ideas Box access on students' psychosocial outcomes in the Danish Refugee Council (DRC) community centre, and (2) an evaluation to determine trends in students' examination scores after fully integrating an LWB-managed Ideas Box into Johud's non-formal education curriculum.

We measured the impact on psychosocial outcomes of introducing the Ideas Box programme into the DRC centre using a combination of difference-in-difference analysis and propensity score matching. For this purpose, we used data collected by the DRC in its community centres in Amman, Karak, and Mafraq. The DRC developed the data collection tools to measure psychosocial outcomes based on a review of validated tools and focus group discussions with students in DRC centres. The questionnaires included 5-point Likert-scale questions on (1) prosocial behaviour, (2) cognitive and emotional functioning, (3) daily tasks and problem-solving, and (4) self-esteem. Possible responses ranged from "completely disagree" to "completely agree." The DRC administered the surveys and assessments twice: before and after students participated in a 6-month cycle of weekly psychosocial programming between June and November 2018. For the impact analyses, we created four outcome measures corresponding to the averages of the responses to each of the four types of psychosocial outcomes.

AIR used a parsimonious model, with a nearest neighbour (NN) matching strategy, because it is the most straightforward matching estimator. In an NN estimation, the individual from the comparison group is chosen as a matching partner for a treated individual who is closest in terms of propensity score, and outcomes are compared for the pair. The propensity score is calculated based on available characteristics of the individual. In this case, study participants were matched on age, nationality, and gender. As there was balance, for the most part, across the distribution of propensity scores (see Exhibit I-3), a comparison group individual was used only once as a match (i.e., without replacement). To reduce the risk of bad matches in situations where the nearest neighbour was far away in terms of propensity scores, a "caliper" of 0.1 was used. This imposes a restriction on what the difference in propensity scores for a matched pair can be (Caliendo & Kopeinig, 2008).

We must exercise caution in interpreting the findings from this method because of various limitations. First, the impact results are only valid if the treatment and comparison groups would have experienced parallel trends in the absence of the programme. Unfortunately, we cannot test this assumption. We were also only able to control for three characteristics in our propensity score matching procedure. This may have adversely affected the matching of treatment and comparison groups.

Exhibit I-3. Balance in Observable Characteristics



AIR®

AMERICAN INSTITUTES FOR RESEARCH®

Established in 1946, the American Institutes for Research (AIR) is an independent, nonpartisan, not-for-profit organization that conducts behavioral and social science research on important social issues and delivers technical assistance, both domestically and internationally, in the areas of education, health, and workforce productivity.

MAKING RESEARCH **RELEVANT**

AMERICAN INSTITUTES FOR RESEARCH
1000 Thomas Jefferson Street NW
Washington, DC 20007-3835 | 202.403.5000
www.air.org

LOCATIONS

Domestic: Washington, DC (HQ) | Monterey, Sacramento, and San Mateo, CA | Atlanta, GA | Honolulu, HI | Chicago and Naperville, IL Indianapolis, IN | Metairie, LA | Waltham, MA | Frederick and Rockville, MD | Chapel Hill, NC | New York, NY | Columbus, OH | Cayce, SC Austin, TX | Arlington and Reston, VA | Seattle, WA

International: Algeria | Ethiopia | Germany | Haiti | Zambia

9691_10/19