

ANNALS OF THE NEW YORK ACADEMY OF SCIENCES

Special Issue: *Implementation Research and Practice for Early Childhood Development*

REVIEW

Measuring the implementation of early childhood development programs

Frances E. Aboud¹ and Elizabeth L. Prado²¹Department of Psychology, McGill University, Montreal, Quebec, Canada. ²Department of Nutrition, University of California, Davis, Davis, California

Address for correspondence: Frances E. Aboud, Department of Psychology, McGill University, Montreal, QC H3A 1G1, Canada. frances.aboud@mcgill.ca

In this paper we describe ways to measure variables of interest when evaluating the implementation of a program to improve early childhood development (ECD). The variables apply to programs delivered to parents in group sessions and home or clinic visits, as well as in early group care for children. Measurements for four categories of variables are included: training and assessment of delivery agents and supervisors; program features such as quality of delivery, reach, and dosage; recipients' acceptance and enactment; and stakeholders' engagement. Quantitative and qualitative methods are described, along with when measures might be taken throughout the processes of planning, preparing, and implementing. A few standard measures are available, along with others that researchers can select and modify according to their goals. Descriptions of measures include who might collect the information, from whom, and when, along with how information might be analyzed and findings used. By converging on a set of common methods to measure implementation variables, investigators can work toward improving programs, identifying gaps that impede the scalability and sustainability of programs, and, over time, ascertain program features that lead to successful outcomes.

Keywords: implementation research; measurement; quality; fidelity; nurturing care

Introduction

Documenting the process of implementing a nurturing care, health, or education program is critical for a number of reasons. One is simply to know what was actually implemented so that outcomes can be related to the actual, rather than intended, program. Another is to improve the program if gaps are found. A third is to assist in the scale-up of the program or its adaptation and use elsewhere. Finally, when sufficient comparable information is available from many programs, a multiple regression analysis can be conducted to reveal which implementation components lead to better outcomes. Good measures and methods of measurement are important in order to arrive at credible conclusions. The purpose of this paper is to outline how to measure implementation of an early childhood development (ECD) program.

Methods might differ depending on the goal of the implementation. If the goal is to conduct a pilot of the intervention in order to determine its feasibility in the adapted form and identify problems in delivery and demand, then variables related to delivery agents (also known as service providers) and recipients might be foremost. This information would be useful to further adapt the program's curriculum and delivery to the context, create better training for providers, and mobilize demand. In contrast, if the goal is to examine a well-developed program, then the quality of the program and its fidelity to what was intended may be equally relevant. Most researchers would wait until they have a solid program before evaluating outcomes along with the implementation process.

We first address some overarching issues regarding the measurement of implementation, such as conceptual frameworks, standardization versus

doi: 10.1111/nyas.13642

Ann. N.Y. Acad. Sci. 1419 (2018) 249–263 © 2018 The Authors. *Annals of the New York Academy of Sciences* published by Wiley Periodicals, Inc. on behalf of New York Academy of Sciences.

249

This is an open access article under the terms of the Creative Commons Attribution License, which permits use, distribution and reproduction in any medium, provided the original work is properly cited.

tailoring to your context, using independent or stakeholder data collectors, balancing fidelity and flexibility, use of mixed methods, and how to document the context and other inputs. The main body of the paper is a practical description of how to measure a number of outputs and immediate outcomes that are part of the implementation process. Finally, we end with an outline of how to analyze and use the information for specified goals such as improvement, scale-up, and identifying key features of effective programs.

Overarching issues

Implementation research when used to describe and evaluate the way an intervention is being carried out may be seen as an extension of monitoring, or the strict reporting of activities intended to occur while rolling out a program. It is often also seen as a part of the evaluation of impact in that knowing what was done helps explain its effect. However, we consider implementation research as the systematic collection of information on how a program or intervention is carried out and contextual factors that bear on it.¹ It goes beyond simply reporting on intended activities and coverage (monitoring) to include whether delivery meets current standards of quality and whether providers have been trained to standards of competence. It may include solutions to improve the program and to promote scale-up and sustainability.

Conceptual frameworks

Different conceptual frameworks have been offered by authors to organize the process and the variables involved in implementation research. For example, Duncan *et al.*² use a timeline in their framework by specifying that there are initial considerations regarding the setting, therefore one must plan the implementation research and then conduct it. Others outline variables to be assessed: fidelity, dosage, quality, participant responsiveness, program differentiation, reach, and adaptation.³ Durlak and DuPre⁴ also include a number of features of the community, provider, and organization, which, as stakeholders, can have a facilitating or hindering effect on implementation. Peters *et al.*¹ add acceptability, feasibility, cost, and sustainability to the list.

Our framework relies on the timeline proposed by a logic model of ECD, where inputs, outputs, and immediate outcomes are the crux of an implementation process that allows others to see the pathway of change driving outcomes (see Fig. 1). The variables we fit to these different layers come mainly from the fidelity framework of Borrelli *et al.*⁵ and a set of reporting guidelines for implementation of ECD programs recently proposed by Yousafzai *et al.*⁶ Information about inputs should be documented at the planning stage. This includes information about the societal and community contexts, the organizational capacity of those providing services, and program resources.⁴ This information may come

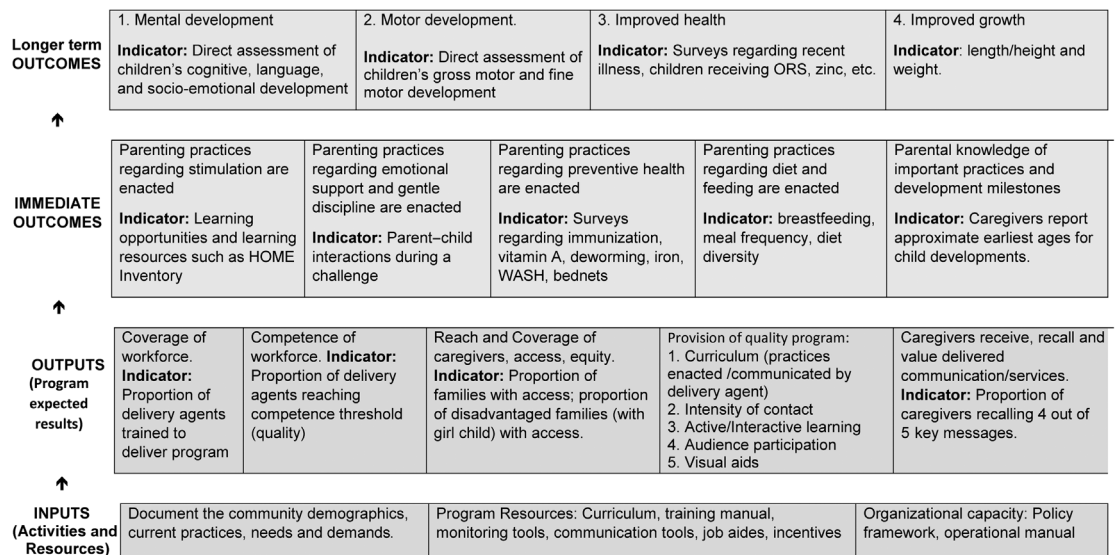


Figure 1. Logic model depicting conceptual frame and pathways for attaining goals.

from existing data sources and formative qualitative methods. In our discussion here, we focus mostly on a set of critical variables from the outputs and immediate outcome levels of the logic model (Fig. 1), including how delivery agents were trained, how the program was delivered, how stakeholders were engaged, and how recipients accepted the messages. We describe what construct is to be measured, how to measure it, who collects the data, from whom, and when. We provide examples from ECD programs, as well as child health, nutrition, and education, where relevant; one particularly complete example is the implementation research of an infant and young child feeding program in Bangladesh, beginning with formative research on the context, the implementation process, scale-up, and sustainability.⁷ Our discussion excludes longer term outcome variables, such as direct measures of children's mental and motor development, because a description of methods for their measurement can be found elsewhere.⁸

Standardization versus tailoring to context

Implementation measures, in contrast to outcome measures, are less standardized; they are necessarily tailored to a program. However, a core or generic set of items found in implementation, as described below, are useful to the extent that content and outputs are similar across programs. In general, it is more useful to start with a core set of previously used items and modify them for a purpose than to develop an entirely new set. It is also important to have individuals who are independent of the implementation collect the information, as this avoids bias arising from conflict of interest. This does not mean that professional researchers need to collect data, but that data collectors be trained to a standard of objectivity. Obviously, there is an advantage to having stakeholders such as supervisors, delivery agents, and even recipients observe and rate the quality of the delivered program on different occasions, as this enhances their engagement and capacity to understand quality. In this case, strategies to reduce bias and enhance validity are important. The more comparable different measures are across studies, the greater the opportunity for an integrated analysis of findings, and, hence, the more convincing the conclusions will be about what implementation features lead to effective outcomes.

Fidelity and flexibility

Tensions between fidelity or adherence to the intended program and adaptation/flexibility will arise. Certain core features of the program need to be identified and observed when delivered. Features such as messages about responsive stimulation and use of gentle discipline, along with an active learning method of behavior change, are considered core features of parenting programs. Likewise, free-choice opportunities for indoor play with a wide selection of play materials and playmates are core features of most preschools. Core features should be those that have some evidence supporting their link to desired outcomes, such as nurturing care practices and child development. Adaptation to a context calls for some modification, but rarely in core features.^{9–11} For example, changes in emphasis on nutrition messages may occur depending on the need in that context. One place may require an emphasis on animal-source foods, while another place may require an emphasis on responsive feeding. Need in each context might be informed by past research or a current survey. Cultural adaptation of illustrations, vocabulary, playthings, songs, and stories can usually be done by the local providers. Delivery agents will need to be informed about core features and about other features that can be modified, such as flexible time allocations for different activities and flexible reviews of past activities that need more attention. Individual recipients with different needs require flexible attention. In brief, all stakeholders need to comment on the feasibility and acceptability of the program for a given context, and delivery agents need to know the fine line between what is core and what is flexible when implementing sessions. If, on the basis of early implementation feedback, problematic features of the program are identified, then changes should be systematically written into the program and implemented thereafter.

Mixing methods

The measures may use quantitative and/or qualitative methodologies. By using a combination of both, one can derive summary scores and also vivid descriptions and subjective perspectives of participants.^{12,13} The use of quantitative and qualitative methods in sequence allows one to specifically ask respondents to comment on the meaning of the quantitative finding. For example, if delivery agents

are performing poorly in a competency test, they can be asked why, and what factors are interfering with their learning and/or performance. Similarly, if caregivers do not increase the stimulation they provide to their children, they can be asked what kinds of barriers they encounter (internal as well as external) and how enablers can be increased. When different methods are used to tap a variable, researchers often attempt to integrate findings, or check one against another in order to validate them. However, a lack of convergence can be equally interesting; for example, one might find that supervisors, delivery agents, and recipients express different opinions as a function of their different perspectives. Managing these different perspectives is critical to success of a program, particularly by addressing concerns of the recipient.

Documenting inputs and context

Although not necessarily measured, implementation reports should document their context and inputs. This includes information about the national and community contexts, the organizational capacity of those taking on service provision, and program resources. Information about context might include past need or demand for nurturing care and child development from demographic and health surveys or other reports, along with documentation of existing and past programs. As part of preparation for an ECD program, for example, one would conduct a survey of current parenting practices, use of health services, and education levels of parents. Such formative research might also include qualitative interviews of parents and other stakeholders. Government policy, local expertise, training in ECD, and activities of advocacy groups and stakeholders would help inform about capacity and acceptability.¹⁴ Adaptation of the program should take into account such information. Finally, it is important to document the program itself: its curriculum and method of teaching/learning, training manuals, and measurement tools.

Review of measures and methods

Our examples of measures come from the two main types of early childhood programs, namely, those delivered *directly* to children, as in group care or preschool settings, and those delivered to their caregivers, such as parents, and thereby *indirectly* to children.

When the objective is to describe and evaluate the implementation of a preschool program, it is important to plan for the assessment of preschool teachers' expertise and how they are supervised, for an independent observation of the quality of the teaching/learning setting, for enrollment and attendance of girls and boys along with their socioeconomic background, and for the participation of parents. Engagement and commitment on the part of the responsible ministry of education are also key. If the preschool program has already been in existence for a few years, then the timing of these assessments may be mid-year; if it is a new program, then assessments may be interspersed several times throughout the year, so that problems can be detected and repaired early. One study found, for example, that additional teacher training helped to improve the quality of a Colombian preschool,¹⁵ while another study in Indonesia required a review of policies governing teacher training and contact time with students.¹⁶

When the objective is to describe and evaluate a parenting program for children under 3 years, the plan might include an assessment of the delivery agents' abilities to communicate accurately and engagingly with parents, an independent observation of the fidelity of the sessions with caregivers, and a description of how caregivers are receiving the messages and enacting them at home. The last, namely caregivers' reception and enactment of the practices, is a key feature of parenting programs that, to be effective, rely heavily on caregivers' behavior change. Because well-established models of such programs do not yet exist, the barriers and enablers encountered by caregivers, delivery agents, and program managers need to be explored. If fidelity is poor and caregivers are not enacting the desired practices, then barriers and enablers will need to be scrutinized in order to improve the program. Once again, if the program is new, the plan should allow for several interspersed data collections in order to make timely improvements to the program. Supervisors, program managers, and delivery agents might be involved with an independent researcher to provide information, as in the Pakistan Care for Child Development program described by Yousafzai and colleagues.¹⁷

Below, discussion of measures and methods is organized into sections dealing with delivery agents and supervisors, program features, recipients, and stakeholders. Figure 2 summarizes measures and

Variable	Method	Program Phase		
		Planning	Preparation for Implementation	Implementation
1. Delivery Agents (Service providers)				
Training delivery agents and assessment of competencies to deliver the program	Quantitative	<ul style="list-style-type: none"> Competency tools Standardized role-playing scenarios 	<ul style="list-style-type: none"> Number of delivery agents trained Hours of training received Education level of delivery agents Knowledge-based evaluation (e.g., written test score) Practice-based evaluation (e.g., checklist or rating scale) 	<ul style="list-style-type: none"> Hours/frequency of ongoing training/coaching received Knowledge-based evaluation (e.g., written test score) Practice-based evaluation (e.g., checklist or rating scale) Percentage of delivery agents who passed evaluations (e.g., scored 80% or above)
	Qualitative	<ul style="list-style-type: none"> Description of training methods and materials 	<ul style="list-style-type: none"> Trainee evaluations of what was helpful and what could be improved in the training 	<ul style="list-style-type: none"> Delivery agent log of new skills learned and applied
Training supervisors and assessment of competencies to train, monitor and mentor delivery agents	Quantitative	<ul style="list-style-type: none"> Competency tools Standardized role-playing scenarios 	<ul style="list-style-type: none"> Methods for delivery agents listed above can also be applied to supervisors (number of supervisors trained, education level of supervisors, etc.) 	<ul style="list-style-type: none"> Hours/frequency of ongoing training/coaching received
	Qualitative	<ul style="list-style-type: none"> Description of training methods and materials 	<ul style="list-style-type: none"> Methods for delivery agents listed above can also be applied to supervisors 	<ul style="list-style-type: none"> Supervisor log of new skills learned and applied
2. Program				
Quality of delivered session	Quantitative	<ul style="list-style-type: none"> For preschools, adapt standardized tools such as the MELE Review existing parenting tools and modify for your needs 	<ul style="list-style-type: none"> Reliability of supervisor compared to an independent observer observing and rating parenting and early learning settings 	<ul style="list-style-type: none"> Checklist or systematic set of ratings (e.g., MELE) completed by an independent observer or supervisor
	Qualitative	<ul style="list-style-type: none"> During piloting of program, interviews with all stakeholders 		<ul style="list-style-type: none"> Interviews with the delivery agent, supervisor, and recipient probing what was done and why
Reach and attendance	Quantitative		<ul style="list-style-type: none"> Numbers of disadvantaged parents with an age-eligible child. Community's and parents' interest in attending parenting sessions or sending child to preschool. 	<ul style="list-style-type: none"> Percentage of scheduled sessions that were delivered overall and by each delivery agent Percentage of eligible or targeted individuals who participated in at least one session, at least half of sessions, at least 80% of sessions, etc.
	Qualitative			<ul style="list-style-type: none"> Delivery agent and supervisor evaluations of how active and engaged were participants. Interviews with the delivery agent, supervisor, and recipient probing barriers to attendance and participation
3. Recipients				
Recipients' acceptance, recall and enactment of program content	Quantitative	<ul style="list-style-type: none"> Adapt standardized tools such as the HOME Inventory or Observation of Mother and Child Interaction 	<ul style="list-style-type: none"> Situation analysis of need, e.g., education level of recipients, currently enacted parenting practices, net enrollment ratios for preprimary and primary. 	<ul style="list-style-type: none"> Percentage of key messages caregivers are able to recall Percentage of targeted behaviors caregivers report or are observed to enact Percentage of content-related questions caregivers correctly answer Percentage of caregivers who reached a threshold of behaviors/correct answers (e.g., 80% or above) Standardized tools such as the HOME Inventory or Observation of Mother and Child Interaction
	Qualitative		<ul style="list-style-type: none"> Cultural taboos and norms regarding child rearing (relevant to adaptation of program content and method). 	<ul style="list-style-type: none"> Exit survey with caregivers after a session, asking if they are satisfied with the session, trusted the delivery agent's advice, felt supported by the delivery agent and by other attendees, and promised to follow up by enacting the practices at home In-depth interviews to explore barriers and enablers to enacting practices

Figure 2. Measurement of outputs and immediate outcomes during three program phases.

4. Stakeholders (including organization personnel, policy makers)

Engagement of Stakeholders	Quantitative	<ul style="list-style-type: none"> • Number and categories of stakeholders (e.g., government, providing organization) • Schedule for meetings 		<ul style="list-style-type: none"> • Frequency of stakeholder meetings • Decisions made by stakeholders • Evidence of acceptability and adoption by stakeholders
	Qualitative	<ul style="list-style-type: none"> • Description of the role played by government and non-government stakeholders 	<ul style="list-style-type: none"> • Prepared guidelines for focus group discussions with stakeholders regarding engagement with program, capacity building. 	<ul style="list-style-type: none"> • Description of the role played by government and nongovernment stakeholders; leadership, capacity • Meeting minutes, with content analysis of discussion and decisions made • Focus group discussions with delivery agents about their acceptance of the messages they are to deliver, their feeling of connection with caregivers, and their belief in the value of the program
Barriers and enablers of program delivery and receipt from the perspective of all participants	Qualitative		<ul style="list-style-type: none"> • Prepared guidelines for focus group discussions and some individual interviews with stakeholders regarding barriers and enablers to scale and sustainability. 	<ul style="list-style-type: none"> • Focus group discussions with groups of stakeholders. • Individual interviews to allow expression of personal or sensitive barriers

Note. Some data collected using a qualitative method may be coded and quantified for statistical analysis.

Figure 2. Continued.

indicators for these variables, and provides a timeline for their measurement.

Delivery agents and supervisors

Training delivery agents and assessment of competencies to deliver the program. Training manuals outline how delivery agents are to become competent at delivering the program. Most will be unfamiliar with the content of the program, for example, the practice of providing psychosocial stimulation and talking with an infant, and with the active/interactive behavior-change format. Desirable competencies include knowledge of how to provide stimulation, demonstration of the practice itself, communication to motivate recipient(s), and engagement of recipient(s). Role plays with other trainee delivery agents may be one of the strategies used to both train and test the competencies of delivery agents. The number of strategies used will depend on the prior experience and expertise of the individual delivery agent. So, recruitment of personnel and their desired levels of education and experience should be reported along with the method of training and testing. If delivery agents are given only general principles and must put together a lesson plan of what to say and do with recipient(s), then lesson development would be another competency to evaluate.

Competencies that are knowledge based can be assessed with a paper-and-pencil test, whereas those that are practice and communication based will

require actions be assessed. Role plays have often been used for this purpose, either with simulated recipients or real caregivers and their children. The number of people trained, men and women, and the quality of their performance are also important indicators.

As an example, a standardized test was used to assess the competencies of mental health paraprofessionals in a Nepali community-based program.¹⁸ Because insufficiently trained paraprofessionals can exacerbate rather than relieve the symptoms of mentally ill people, their competence needed to be reliably assessed. Items were initially generated to assess paraprofessionals' competencies in specific skills, such as behavioral activation, and general skills, such as therapeutic alliance and empathy. Only items that provided reliable answers from different professionals observing actual interactions or videos of interactions were included in the test. After training, paraprofessionals engaged in role plays with standardized patients and then answered questions about their judgments of the case. As another example, a program in Lombok, Indonesia used a specific method to test community health workers who were to deliver a nutrition program to pregnant women.¹⁹ Not only did they have to pass a sequence of tests as they progressed through training, but also their knowledge, social skills, and practical duties were subject to a "head, heart, and hands" assessment. The head score was based on assessment of knowledge retained during trainings. The

heart score was based on how well the community worker demonstrated caring for the pregnant woman and her health, measured by inputs from the pregnant woman and the supervisor. The hands score reflected how thoroughly and efficiently the fieldwork assigned to the health worker was completed. Those rated higher were more effective at reducing early infant mortality among their clients.

Training preschool teachers is now part of certification programs in places like Kenya, where teachers receive courses on pedagogy and ECD, along with on-the-job coaching and assessment. An intensive assessment of competency was used in two published studies of preschool teachers. One in the United States and one in Chile attempted to improve teachers' skills at emotional support of children, along with instructional support in teaching language and literacy.^{20,21} In their description of the implementation process, researchers show how trainers conducted monthly observations and teachers kept weekly logs of their new skills. An observation of the teachers in action in their classrooms showed improvements on both skill sets, but particularly emotional support.

Training supervisors and assessment of competencies to train, monitor, and mentor delivery agents. Supervisors may need training on the program itself and should be involved as trainers of the delivery agents whom they are to supervise. They will need to be included when the delivery agents' training manual is developed and when the program manual is adapted or modified. Supervisors should be selected on the basis of their higher level of education, experience, and expertise in ECD compared with delivery agents. These can be measured by competency tests in ECD, along with observation of their skills at delivering the sessions (similar to assessing skills of delivery agents). To be credible to delivery agents, supervisors will need to exhibit skills at training of delivery agents and of actually delivering the program. They will need to have a broader and deeper grasp of ECD than the delivery agents they supervise.

Supervisors need to be trained in and tested on supervisory skills, as they are the main interface between the program manager and the delivery team. It is important to record how supervisors are trained to conduct monitoring of sessions, home visits, or early group care settings; they will need to

provide consistent feedback on the delivery agents' performance with the help of a standard monitoring form; and they will need to train and mentor for improvement. An operations manual outlines how frequently training and retraining occurs, how often monitoring takes place, and what other inputs are needed, such as job aids, arranging for peer-to-peer support, and community mobilization. Implementation researchers will want to record if these actually took place.

Yousafzai *et al.*¹⁷ found in Pakistan that the help provided by supervisors changed somewhat over time: as delivery agents became more confident and proficient at counseling best practices, they needed more help on solving caregiver problems, integrating nutrition and stimulation practices, and using more varied methods to demonstrate and explain the practices. When delivery agents were asked to comment on barriers and enablers of the supervisory process, they appreciated supervisors' positive attitudes toward coaching and correcting errors in delivery. This emphasizes that supervisors need to be flexible and attentive counselors themselves, and always one step ahead of the delivery agents they are coaching.

Program quality, coverage, and dosage

These implementation features are often considered "fidelity" or how closely the actual program delivered to the recipient (caregiver and child) matches the activities and dosage intended by the designed program.

Quality of delivered session. Quality refers to features of the program that are the evidence-based core of content and method of teaching/learning. Regardless of the match between the delivered and designed program, if the messages and methods of delivery are inappropriate, quality will be low. Regardless of whether the contact session is with a group of caregivers, with an individual family during a home or clinic visit, or in a group care setting, evaluation of the quality of delivery should entail observations at a minimum. The observer should complete a checklist or standardized set of ratings to derive a quantitative score. They may also include interviews with the delivery agent, supervisor, and recipients; these provide different perspectives on what was done and why.

Observations would normally be conducted early in the program, during a pilot/preparation phase,

and on several later occasions. After examining the quality of early sessions, program designers and other stakeholders might decide to modify the intended program—its content, delivery format, engagement of recipients, spacing of sessions, and duration of sessions. For example, if the delivery format as intended is found to be too didactic and disengaging for the audience, more practice-based activities might be introduced and more discussion generated among recipients.

It is often convincing to have an independent observer, but many programs may ask supervisors or other program personnel to conduct observations and interviews. An advantage of an independent observer is that such a person may feel less social pressure to give a positive review that puts themselves and the delivery agent in a good light. On the other hand, an advantage of a supervisor is that such a person can enhance a positive and ongoing relationship with the delivery agent by providing coaching, feedback, and problem-solving support. Observations based on easily observable behaviors in the setting do not require a professional with expertise in ECD; most researchers hire local university graduates from the social sciences to observe and rate delivery quality and give them adequate training to reach a standard of accuracy and reliability.

How might the quality of a group session with caregivers be observed and rated? A list of important qualities could be generated based on the essential features of the session. A short list of 10–20 critical items, rated with a binary yes–no format, would help ensure that it was used and valued.^{17,22} For example, one item might be that accurate content is delivered; another might be that caregivers have an opportunity to directly interact with their child using the desirable practice. Items relating to caregiver participation may include whether they are given the opportunity to talk to each other, offer support, and encouraged to raise any concerns with the delivery agent. If caregivers are given the opportunity to participate but do not, then this should be noted, as it reveals a problem in the delivery agent's ability to encourage participation. Summary ratings are created to ensure that a minimum standard is achieved. Ratings for an individual delivery agent on a particular day are also discussed with the delivery agent in question in order to improve performance.

The quality of a home or clinic visit with a single family might be based on similar features. More

emphasis might be given to engaging the caregiver, encouraging interactions with the child, praising and coaching the caregiver, asking about barriers and how to solve them, and directing personal attention to the caregiver and her/his problems. Directly picking up the child or interacting with the child in place of the caregiver may be considered too intrusive; but praising the child, in order to model certain behaviors, might be appropriate. Good examples can be found in the Reach Up home visiting program delivered in Zimbabwe¹¹ and the Getting Ready home visiting program delivered in the United States.²³

The quality of a child care or preschool setting requires observing different features and a different delivery format, compared with a parenting group or home visit. In former case, one or more caregivers meet regularly with a group of children from other families and directly interact(S) with them. The quality of these programs, both the setting and the caregivers' training, has been the subject of many observational measures.^{24,25} This is one clear case when the observed quality is not compared with the intended program but rather with what is considered "standard" quality. Ministries of Education may state that they are excluding indoor play for preschool children, but regardless of their intentions, indoor play is considered critical for cognitive and social development of young children. In brief, qualities such as a safe and hygienic environment, responsive adult–child interactions, language experiences such as story reading, opportunities to learn math concepts, and sufficient time and materials for creative and constructive indoor play are present in most measures. In high-income and even middle-income Latin American countries, commonly used quality measures were the Infant/Toddler Environment Rating Scale (ITERS-R) and the Early Childhood Environment Rating Scale (ECERS-R).^{26,27} For example, the ITERS has been used in Ecuador and other Latin American countries, such as Bolivia, Chile, Colombia, and Peru, yielding low quality ratings on physical and also on process features. The ECERS-R was used in Colombia to evaluate the successful improvement of aeioTU preschools for children 3–6 years.¹⁵ Measures to replace the ECERS are now available for preschools in low- and middle-income countries (LMICs). Specifically, the Measure of Early Learning Environment (MELE) by the MELQO Initiative is a quality

measure that was designed to evaluate preschools in LMICs.²⁸ It has been used in a variety of countries in East Africa, East Asia, and Latin America with a core set of items as well as ones modified for the context.

The quality of a program is critical and should be measured with care. It indicates not only whether delivery agents are performing as expected, but also more importantly whether the program is of sufficient quality to result in benefits to children's development. The advantage of using a quantitative quality measure is that quality can be correlated with child learning and development outcomes. This has been done with preschool quality.^{16,29–31} One study spanning several years found that preschool quality was initially low and that children were not performing better than a control group.³⁰ Only after several years of improving the play and teaching/learning activities did students perform better.

Coverage and dosage (reach and attendance).

Most programs target a certain audience. Its reach, then, refers to what proportion actually participated. Beyond this, most programs keep attendance records for recipients as well as the number of sessions or visits attempted by the delivery agent. How often did the family attend the clinic with their child, how often was the caregiver at home for the expected visit, and how often did caregivers attend the group session? How much of the curriculum was actually delivered by the agent? This is sometimes referred to as dose delivered and dose received. The number of sessions attended may be used as a modifier of the outcome, comparing high- versus low-attenders. Attendance records may also reveal how much demand there is for this service. Although most early childhood reports include attendance at sessions and compliance in offering the nutrition supplement,³² it is usually a single summary score rather than a more helpful range, such as how many attended 50–75% and how many 75% and over. It is also informative to know who maintained consistent attendance and who did not. One study found that more educated mothers attended few sessions but attained the same knowledge outcome score.³³ This is informative regarding the literacy- or education-level of the group program, an issue that would be solved with tailored messages in individual home or clinic visits.

Fathers may also be reluctant to attend parenting sessions because they expect mothers to play a more important role in nurturing care. However, new strategies are being tried to entice fathers to learn about caring for their own children. Some strategies entail addressing marital issues such as conflict and communication that also include discipline of children.²² So, attendance by fathers is worth measuring and reporting along with the strategies for encouraging attendance. Attendance by grandmothers and adolescents, if they are invited, is also a new feature of some programs.³⁴ Overall, the reach of the program is important to report, namely, how many people had access to the program and what proportion actually participated.

Recipients' acceptance, recall, and enactment of program content

If a program is directed at caregivers who are then expected to interact in specific ways with their children, it is advisable to find out if they are accepting, attending to, recalling, and enacting the desirable practices with their child. If they are not, then child outcomes are unlikely to improve. Self-report and observation are the methods of choice because the data must be collected from individuals. Some programs conduct an exit survey with caregivers after a session, asking if they are satisfied with the session, trusted the delivery agent's advice, felt supported by the delivery agent and by other attendees, and promised to follow up by enacting the practices at home.¹⁴

A midline interview with a subsample of caregivers could clarify if they have retained and enacted the practices being advocated. Again, an independent interviewer would provide more credible information, but a program supervisor if well trained could do the same. It is important to minimize asking loaded questions or arousing a socially desirable response on the part of the recipient. A minimal question such as, Do you remember any advice that you were given on how to care for your child?, can be followed by "What else?" each time the caregiver responds, until no more can be added. Vague answers such as "feed a balanced diet" should be probed with "What is a balanced diet?" Spontaneous recall (e.g., "Do you remember any advice . . .?") is better than recognition recall (e.g., "Did they tell you about how to talk with your child?") but

spontaneous recall always requires probing with “What else?”

The enactment information can be elicited with a question about each recalled and then nonrecalled practice. Questions phrased nonjudgmentally such as “Are you able to do any of this?” encourage caregivers to provide a nuanced answer detailing how often and in what ways they can do it. For example, the message “Talk with your child” may be enacted “only sometimes when I am alone with the child and not pre-occupied with cooking or cleaning; sometimes I even forget, but my neighbor reminds me.” Scoring these answers requires a graded scale from, say, 0 to 5, taking into account frequency, situations, and motivation. The scoring must show inter-rater reliability.^{22,35} If the analyses show that on average only two or three messages out of five are recalled, then this suggests that messages are not salient, not being learned in a personal or active manner. If enactment scores are on average 2 out of 5, then caregivers are having trouble doing them eagerly on a daily basis. Inquiring about barriers and enablers to performance should follow: What makes it easy and what makes it hard to do? Barriers and enablers may be first coded into categories relating to individual (internal), interpersonal and structural features (external), and then simply counted per respondent.^{22,35} The data may also indicate that some people are performing the practices regularly and others are struggling. The program needs revision, perhaps by offering home visits to those who are struggling, or by engaging everyone in a discussion of barriers and how to overcome them.

A more structured way to determine whether psychosocial stimulation practices are being enacted by caregivers is with the 45-item Home Observation for Measurement of the Environment (HOME) Inventory.³⁶ It measures the quantity and quality of support and stimulation provided to the child at home. Some of the items require the observation of how a caregiver interacts with the child during the interview, and other items require a self-report (e.g., regarding trips to the market). It has been modified to describe comparably challenging playthings available to children in LMIC, and generally requires the data collector to observe them. A version is available for children 0–3 years and another for those 4 and 5 years who are expected to have more stimulation in shapes, drawing, and books. Although the 45-item measure taps six domains of

support and stimulation, factor analyses rarely verify this structure with LMIC data; the total score provides a reliable summary. A 10-item short version, called the Family Care Indicator, is solely a caregiver report asking about stimulating materials and activities provided to the child. It is appropriate for a wide age range of birth to 5 years, but not for a narrow age range of birth to 2 years, as is often the case in nurturing care programs. Some items have been found to asymptote at 5 months, while other items such as reading remain low across ages.^{37,38} The HOME is often used at the beginning and end of a program to examine benefits to the intervention group. Because the HOME Inventory is such a strong determinant of mental development in children, raising this score is an important objective of psychosocial stimulation interventions. If scores do not advance sufficiently, then a more robust program is required. A more detailed examination of how caregivers talk with their child in a responsive manner is measured directly with the Parent-Child Picture Talk measure, standardized as Observation of Mother and Child Interaction.^{39,40} Noting the frequency of a number of cognitive- and language-supporting remarks made by the caregiver reveals the extent to which verbal stimulation is being provided.

Some implementation programs focus on building caregivers’ knowledge about young children’s capabilities at different ages and stages. They would therefore want to know whether caregivers’ knowledge improved with input from the delivery agents or from visual aids and brochures. Some milestone knowledge questions focus on the cognitive capabilities of an infant, given that these are less obvious or visible to a caregiver, but of crucial importance. Questions should be about generic child development and not an assessment of any particular child; parents should not be put in the position of evaluating their own child’s development. Questions may follow the format: At what age is a child able to (1) recognize the mother, (2) enjoy seeing colorful moving objects, (3) be interested in hearing someone talk, (4) understand some spoken words, (5) smile when excited, (6) use gestures and sounds to communicate needs, and (7) learn to do something by himself/herself, such as putting stones in a cup. If caregivers do not know that these capabilities are present well before the child’s first birthday, they will not know why they are being asked to provide stimulation from birth.^{41,42}

Engagement of stakeholders

Stakeholders include all those organizations and groups of people who have an interest or involvement in the program, such as policy makers, managers, providers, and recipients. The most critical group might be the organization or government ministry responsible for funding and initiating the program. The role they played in initiating the program, in providing resources, in selecting the content and delivery format, and in supporting it with advocacy and a policy framework is useful information for sustainability. So are the partnerships they formed with others to expand reach, expertise, and resources for early childhood programs. Sometimes, the method of data collection for a stakeholder is simply minutes of meetings along with a content analysis of comments and decisions made at such meetings. Frequent meetings might take place at the initial stages of the program, fewer during its delivery, and again occurring frequently at the end as findings are disseminated and plans made for sustainability and advocacy.

Another important stakeholder is the organization responsible for day-to-day management of the program. Their role in developing the content or adapting for the context, as well as developing an operational framework, and managing human and other resources would be useful information. This group might have different perspectives, so it would be reasonable to interview some top management staff privately and others in focus groups. One objective would be to find out what procedures were easy to incorporate into the existing structure and what procedures required quick learning or new solutions. Were certain aspects of the program altered because of management issues, for example, the number of days of training allowed, or the qualification level of delivery agents? How often did managers meet with others involved in implementing the program so that all felt they could express concerns and get them resolved?

Delivery agents also need to be engaged if they are to take on the workload, perhaps doing something different from their normal activities. They have to be in agreement with the messages and services they are being asked to deliver. If the new practice is viewed as harmful, there should be a full discussion with the community and a decision made about its inclusion. In the end, caregivers choose

for themselves whether they want to enact it at home. Focus group discussions (FGDs) with delivery agents could inquire about their acceptance of the messages they are to deliver, their feeling of connection with caregivers, and their belief in the value of the program.

Qualitative methods regarding engagement are considered better than surveys or written answers from respondents. Nonetheless, the questions should be explicitly worded to elicit the stakeholders' engagement in the program. Do they feel some personal responsibility for the smooth operation of the team effort, some agreement with the goal and method of the program, and support from peers and superiors? Content analysis of respondents' answers would be more useful than thematic analysis because the categories of answers are already known.⁴³ The goal is to enhance engagement of all stakeholders and their ability to work with each other.

Barriers and enablers of program delivery and receipt from the perspective of stakeholders.

In order to improve the program and understand specific problems people face, it is important to explore with key groups both barriers and enablers. At a minimum, program managers, delivery agents, and recipients should be given the opportunity to express their evaluations. This might be done mid-stream or earlier if problems arise that clearly constrain implementation; it should also be done at the end to give some closure. The method of choice is a FGD with 8–10 people at a time, and individual interviews could be used to provide a safe place to express personal or sensitive barriers.²² The persons conducting FGDs with program managers and delivery agents should be those who have no conflict of interest. Some researchers always audio record such discussions, but many respondents will not feel comfortable with this, so one or two people may take notes and also record whether faces in the audience show general agreement or disagreement with comments. The moderator of the discussion groups should have an outline of topics covering all the critical inputs and outputs within their sphere of responsibility and perhaps within the broader program.

Data analysis to meet four goals

Initially, we described four goals for the collection of implementation information. Each requires a

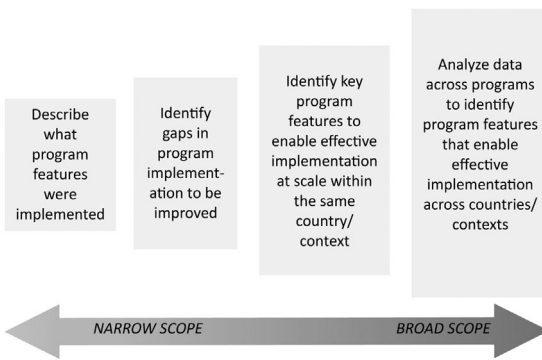


Figure 3. Purposes of implementation research in early childhood development programs.

slightly different analysis so that findings can lead to appropriate decisions that vary in scope relating to one program or many (Fig. 3). The first is simply to know what was actually implemented so that outcomes can be related to the actual rather than the intended program. Descriptive analyses are needed on indicators such as how many delivery agents and supervisors were trained, how long it took to reach the standard, and how frequently they were supervised. Likewise, quality indicators on the provision of the sessions can be analyzed separately for different domains (e.g., content, method, and recipient participation) as well as a summary score. There may be summary scores for different delivery agents across multiple time points. Assessment of recipients' recall and enactment should be less frequent, or should involve contact with different recipients at different times so that they do not feel overburdened. Descriptive statistics on how many messages are recalled and enacted and what barriers and enablers are influential will be important. Several researchers have drawn attention to the need to distinguish between what was provided to the intervention group and what to the comparison group. In some cases, the two groups were found to overlap much more than expected.^{23,44}

For the purpose of identifying gaps in program implementation to improve the program, one might want to focus on analyses of the quality and recipient data. Standards should be prespecified for what quality is expected and how well recipients should be recalling and enacting the practices. These indicators reveal how well the program is performing, and are not an evaluation of people. Analyses by item and by domain would reveal which qualities were

not meeting standards. For example, observational ratings of quality items on a 1–4 scale might be set at a threshold of 2.5, so that any item below this requires attention, either by increasing resources, retraining delivery agents, or redesigning a component of the program. Sometimes, only certain sites yield low scores, pulling the average down, so they receive special attention. Another way to analyze quality data is by delivery agent. If a yes-no checklist is used while observing parenting sessions, one might ask: Over how many sessions did each delivery agent satisfy the 10 or 15 quality items? If 80% of quality items are met, then the agent is doing well, but if 50% then the supervisor needs to provide support and mentoring to those agents. The performance of delivery agents may also be linked to the recall and enactment data of their corresponding recipients. Performance below threshold may also trigger qualitative interviews to be conducted with delivery agents or recipients probing why that behavior might be difficult to implement in that setting.

Once outcome data are known, it is possible to link them with program features such as the quality of communicating messages about certain nurturing practices or preschool language teaching. One might want to know whether higher delivery agent or supervisor competence was associated with positive program outcomes. The first step would be to calculate an overall competence score for each delivery agent, based on all measurements of that agent's competence throughout the duration of the program. The next step is to merge that delivery agent's competence score with the program outcome scores for the children served by that agent. One could then examine whether delivery agent competence scores are associated with child outcome scores. One could examine delivery agent competence scores as an effect modifier of intervention effects through an interaction term between delivery agent competence and intervention group.⁴⁵

For the purpose of identifying key program features to enable effective implementation at scale within the same country/context, other analyses would be conducted. For example, one might want to know how many hours of training are required for delivery agents of varying education levels to achieve competence delivering the intervention. The first step would be to divide the data into groups by delivery agent education level, such as one group of

delivery agents with less than a high school degree, a second group who graduated from high school, and a third group who had professional training. For each of these groups, calculate the competency score after 40 h of training, 60 h of training, and 80 h of training. In order for this type of analysis to be conducted, competency evaluations must be conducted at multiple time points, throughout pre-program training as well as the continuing period of on-the-job training and coaching. This analysis could reveal a threshold for the number of hours of training required to reach an average competency level among each group of delivery agents. This information could then be used to inform what would be required to implement the program on a larger scale.

The fourth goal is a longer term one, namely, identifying program features that enable quality implementation across countries and contexts and lead to positive outcomes. This requires meta-analysis based on sufficient comparable data. Information would be extracted from published studies reporting variables outlined in this paper, such as the education levels of the delivery agents, hours of training, quantity of key program elements implemented, level of stakeholder engagement, and caregiver practices. Programs could be stratified based on delivery platform, such as home visits versus group sessions, and pooled effect sizes on various outcomes could be generated. A meta-regression could also be used to examine the association of these variables with the effect sizes of the programs on various outcomes.⁴⁶ Before this type of meta-analysis can be conducted, it is essential that researchers and program implementers collect and report this information, using reporting guidelines for ECD elaborated by Yousafzai *et al.*⁶

Conclusions

A large number of early childhood programs have published positive outcomes, with only piecemeal reporting of their implementation process. From these and reports in the health, nutrition, and education sector, we have extracted successful measures and methods of measurement. We have also tried to converge on the variables highlighted in the reporting guidelines for ECD. The coming years will be a time to accumulate evidence on which measures and methods are reliable and provide actionable

data for decision making in ECD programs and policy.

Acknowledgments

The contribution of F.E.A. was to write the first draft of the manuscript. E.L.P. revised its intellectual content. Both approved the final version. Figure 1 was prepared with critical input from Aisha K. Yousafzai.

This paper was invited to be published individually and as one of several others as a special issue of *Ann. N.Y. Acad. Sci.* (1419: 1–271, 2018). The special issue was developed and coordinated by Aisha K. Yousafzai, Frances Aboud, Milagros Nores, and Pia Britto with the aim of presenting current evidence and evaluations on implementation processes, and to identify gaps and future research directions to advance effectiveness and scale-up of interventions that promote young children's development. A workshop was held on December 4 and 5, 2017 at, and sponsored by, the New York Academy of Sciences to discuss and develop the content of this paper and the others of the special issue. Funding for open access of this and the other papers of the special issue is gratefully acknowledged from UNICEF and the New Venture Fund.

Competing interests

The authors declare no competing interests.

References

1. Peters, D.H., N.T. Tran & T. Adam. 2013. *Implementation Research in Health: A Practical Guide*. Geneva: World Health Organization.
2. Duncan, C.M., J.A. Durlak & A. Wandersman. 2012. The quality implementation framework: a synthesis of critical steps in the implementation process. *Am. J. Community Psychol.* 50: 462–480.
3. Durlak, J.A. 2010. The importance of doing well in whatever you do: a commentary on the special section, "Implementation research in early childhood education." *Early Child. Res. Q.* 25: 348–352.
4. Durlak, J.A. & E.P. DuPre. 2008. Implementation matters: a review of research on the influence of implementation on program outcomes and the factors affecting implementation. *Am. J. Community Psychol.* 41: 327–350.
5. Borrelli, B.D., D. Sepinwall, D. Ernst, *et al.* 2005. A new tool to assess treatment fidelity and evaluation of treatment fidelity across 10 years of health behavior research. *J. Consult. Clin. Psychol.* 73: 852–860.
6. Yousafzai, A.K., F.E. Aboud, M. Nores & R. Kaur. 2018. Reporting guidelines for implementation research on nurturing care interventions designed to promote early childhood development. *Ann. N.Y. Acad. Sci.* 1419: 26–37.

7. Sanghvi, T., R. Haque, S. Roy, *et al.* 2016. Achieving behaviour change at scale: Alive & Thrive's infant and young child feeding programme in Bangladesh. *Matern. Child Nutr.* **12**(Suppl. 1): 141–154.
8. Fernald, L., E.L. Prado, P. Kariger & A. Raikes. 2017. A toolkit for measuring early childhood development in low- and middle-income countries. The World Bank, Washington, DC.
9. Baker-Henningham, H. 2018. The Irie Classroom Toolbox: developing a violence prevention, preschool teacher training program using evidence, theory, and practice. *Ann. N.Y. Acad. Sci.* **1419**: 179–200.
10. Gladstone, M., J. Phuka, R. Thindwa, *et al.* 2018. Care for Child Development in rural Malawi: a model feasibility and pilot study. *Ann. N.Y. Acad. Sci.* **1419**: 102–119.
11. Smith, J.A., H. Baker-Henningham, A. Brentani, *et al.* 2018. Implementation of Reach Up early childhood parenting program: acceptability, appropriateness, and feasibility in Brazil and Zimbabwe. *Ann. N.Y. Acad. Sci.* **1419**: 120–140.
12. Creswell, J.W., A.C. Klassen, V.L. Plano Clark & K. Clegg Smith. 2011. *Best Practices for Mixed Methods Research in the Health Sciences*. Washington, DC: Institutes of Health.
13. Palinkas, L.A., G.A. Aarons, S. Horwitz, *et al.* 2011. Mixed method designs in implementation research. *Adm. Policy Ment. Health* **38**: 44–53.
14. Engle, P.L., G. Najimudinova & K. Faromuzova. 2011. Care for development in three Central Asian countries. Report of a process evaluation in Tajikistan, Kyrgyz Republic, and Kazakhstan. UNICEF, New York.
15. Nores, M., A. Figueras-Daniel, M.A. Lopez & R. Bernal. 2018. Implementing aeioTU: quality improvement alongside an efficacy study—learning while growing. *Ann. N.Y. Acad. Sci.* **1419**: 201–217.
16. Brinkman, S.A., A. Hasan, H. Jung, *et al.* 2016. The role of preschool quality in promoting child development: evidence from rural Indonesia. World Bank, Washington DC. Policy Research Working Paper 7529.
17. Yousafzai, A.K., M.A. Rasheed & S. Sijal. 2018. Integration of parenting and nutrition interventions in a community health program in Pakistan: an implementation evaluation. *Ann. N.Y. Acad. Sci.* **1419**: 160–178.
18. Kohrt, B.A., M.J.D. Jordans, S. Rai, *et al.* 2015. Therapist competence in global mental health: development of the ENhancing Assessment of Common Therapeutic factors (ENACT) rating scale. *Behav. Res. Ther.* **69**: 11–21.
19. Shankar, A.V., Z. Asrilla, J.K. Kadha, *et al.* 2009. Programmatic effects of a large-scale multiple-micronutrient supplementation trial in Indonesia: using community facilitators as intermediaries for behavior change. *Food Nutr. Bull.* **30**(Suppl.): S207–S214.
20. Domitrovich, C.E., S.D. Gest, D. Jones, *et al.* 2010. Implementation quality: lessons learned in the context of the Head Start REDI trial. *Early Child. Res. Q.* **25**: 284–298.
21. Yoshikawa, H., D. Leyva, C.E. Snow, *et al.* 2015. Experimental impacts of a teacher professional development program in Chile on preschool classroom quality and child outcomes. *Dev. Psychol.* **51**: 309–322.
22. Singla, D.R. & E. Kumbakumba. 2015. The development and implementation of a theory informed, mother–child intervention in rural Uganda. *Soc. Sci. Med.* **147**: 242–251.
23. Knoche, L.L., S.M. Sheridan, C.P. Edwards, *et al.* 2010. Implementation of a relationship-based school readiness intervention: a multidimensional approach to fidelity measurement for early childhood. *Early Child. Res. Q.* **25**: 299–313.
24. Kaul, V. *et al.* 2012. Early Childhood Education Quality Assessment Scale (ECEQAS). Centre for Early Childhood Education and Development (CECED), Ambedkar University, Delhi.
25. Sylva, K., I. Siraj-Blatchford & B. Taggart. 2003. *Assessing Quality in the Early Years: Early Childhood Environment Rating Scales – Extension (ECERS-E): Four Curricular Subscales*. London: Trentham.
26. Harms, T., D. Cryer & R.M. Clifford. 2006. *Infant/Toddler Environment Rating Scale – Revised Edition*. New York: Teachers College Press, Columbia University.
27. Harms, T., R.M. Clifford & D. Cryer. 1998. *Early Childhood Environment Rating Scale, Revised Edition*. New York: Teachers' College Press, Columbia University.
28. MELQO. 2016. Overview of the Measuring Early Learning Quality and Outcomes (MELQO) initiative. Washington, DC. UNESCO.
29. Aboud, F.E., K. Proulx & Z. Asrilla. 2016. An impact evaluation of Plan Indonesia's early childhood program. *Can. J. Public Health* **107**: e366–e372.
30. Aboud, F.E. & K. Hossain. 2011. The impact of preprimary school on primary school achievement in Bangladesh. *Early Child. Res. Q.* **26**: 237–246.
31. Montie, J.E., Z. Xiang & L.J. Schweinhart. 2006. Preschool experience in 10 countries: cognitive and language performance at age 7. *Early Child. Res. Q.* **21**: 313–331.
32. Yousafzai, A.K. & F.E. Aboud. 2014. Review of implementation processes for integrated nutrition and psychosocial stimulation interventions. *Ann. N.Y. Acad. Sci.* **1308**: 33–45.
33. Aboud, F.E. 2007. Evaluation of an early childhood parenting program in rural Bangladesh. *J. Health Popul. Nutr.* **25**: 3–13.
34. Aubel, J. 2012. The role and influence of grandmothers on child nutrition: culturally designated advisors and caregivers. *Matern. Child Nutr.* **8**: 19–35.
35. Affleck, W. & G. Pelto. 2012. Caregivers' responses to an intervention to improve young child feeding behaviors in rural Bangladesh: a mixed method study of the facilitators and barriers to change. *Soc. Sci. Med.* **75**: 651–658.
36. Bradley, R.H. & R.F. Corwyn. 2005. Caring for children around the world: a view from HOME. *Int. J. Behav. Dev.* **29**: 468–478.
37. Bornstein, M.H. & D.L. Putnick. 2012. Cognitive and socioemotional caregiving in developing countries. *Child Dev.* **83**: 46–61.
38. Bornstein, M.H., D.L. Putnick, J.E. Lansford, *et al.* 2015. A developmental analysis of caregiving modalities across infancy in 38 low- and middle-income countries. *Child Dev.* **86**: 1571–1587.

39. Aboud, F.E. & S. Akhter. 2011. A cluster-randomized evaluation of a responsive stimulation and feeding intervention in Bangladesh. *Pediatrics* **127**: e1191–e1197.
40. Rasheed, M.A. & A.K. Yousafzai. 2015. The development and reliability of an observational tool for assessing mother–child interactions in field studies—experience from Pakistan. *Child Care Health Dev.* **41**: 1161–1171.
41. Ertem, I.O., G. Atay, D.G. Dogan, *et al.* 2007. Mothers' knowledge of young child development in a developing country. *Child Care Health Dev.* **33**: 728–737.
42. Aboud, F.E. & T. Alemu. 1995. Nutrition, maternal responsiveness, and mental development of Ethiopian children. *Soc. Sci. Med.* **41**: 725–732.
43. Hsieh, H.-F. & S.E. Shannon. 2005. Three approaches to qualitative content analysis. *Qual. Health Res.* **15**: 1277–1288.
44. Weiss, M.J., H.S. Bloom & T. Brock. 2014. A conceptual framework for studying the sources of variation in program effects. *J. Pol. Anal. Manage.* **33**: 778–808.
45. Dunst, C.J. & C.M. Trivette. 2012. Moderators of the effectiveness of adult learning method practices. *J. Soc. Sci.* **8**: 143–148.
46. Aboud, F.E. & A.K. Yousafzai. 2015. Global health and development in early childhood. *Ann. Rev. Psychol.* **66**: 433–457.