Strategies to maximize generalization from multiple case studies: Lessons from the Africa Routine Immunization System Essentials (ARISE) project

Sangeeta Mookherji and Anne LaFond

Evaluation 2013 19: 284
DOI: 10.1177/1356389013495212

Article

Evaluation
19(3) 284–303

DOI: 10.1177/1356389013495212
evi.sagepub.com

Strategies to maximize generalization from multiple case studies: Lessons from the Africa Routine Immunization System Essentials (ARISE) project

Sangeeta Mookherji
George Washington University, USA

Anne LaFond
John Snow, Inc (JSI), USA

Abstract
This article considers the challenges of generalizability related to case studies, and specifically for the in-depth case studies of the Africa Routine Immunization System Essentials (ARISE) project. The article describes how these challenges were addressed, by developing a Theory of Change to frame case selection strategies, data collection, and analysis, including synthesis of findings across multiple cases. The authors then consider: the importance of grounding generalizability in theory; balancing within-and cross-case analyses for synthesis; and using theory-based case selection, as ways to support generalizability of the case study findings. Multiple case studies should sequence analysis as: 1) within-case analysis; 2) identification of replicated findings and implementation variation across cases; and 3) synthesis across cases, pooling the data. Case selection should be a stand-alone, formative part of case study research. The lessons from the ARISE case studies suggest that these are important ways in which case study methodology can be strengthened.

Keywords
Africa, case study methods, case study selection, case study synthesis, drivers of performance improvement, evaluation, generalizability, immunization

Corresponding author:
Sangeeta Mookherji, 2175 K Street NW, Suite 200, Washington, DC 20037, USA.
Email: smookher@gwu.edu

Introduction

Maximizing generalizability of findings is the ultimate goal of research: to generate results that are meaningful and relevant as far beyond the study's focal 'population' as reasonable. No matter the subject or the methodology, researchers endeavor to maximize the external validity of their results.

Increasingly, social science researchers are being asked to evaluate complex interventions that are often quite distal from the outcomes and impacts of interest (Byrne, 2013; Douthwaite et al., 2003; Rogers, 2008) and to make statements about causality (Harding and Woolcock, 2012). The challenge of making sense of what works, under what conditions, and why, given the presence of one or more complex interventions, ongoing systems changes, and varying contexts, should not be underestimated. In trying to meet this challenge, international development evaluators are increasingly turning to case studies. However, the results from case study evaluations are often questioned by practitioners, policymakers, and researchers alike, who most often raise doubts about how far a case study-based evaluation can be used for learning beyond the cases that were part of the study ('Case Studies in Development Evaluation: Validity, Generalization and Learning. An International Workshop', 2012). A large component of this doubt stems from a lack of confidence in the generalizability, or external validity, of the findings.

Case study research is especially vulnerable to criticism regarding limited generalizability, given the small 'n' that is its key characteristic. To clarify, in this article we will use Seawright and Gerring's definition of case study research: 'the intensive (qualitative or quantitative) analysis of a single unit or a small number of units (the cases), where the researcher's goal is to understand a larger class of similar units (a population of cases)' (Seawright and Gerring, 2008). Researchers grounded in the concepts of experimental and quasi-experimental designs consider case study to have limited potential for generalization, stating that case study research has problems with representation and inference because of the small size and non-randomly selected sample that is the source of data (Marchal et al., 2010; White and Phillips, 2012; Yin, 2009). However, these criticisms are based on statistical criteria for external validity. Regardless of the principal research approach, or mix of approaches, all research studies, including case studies and 'gold-standard' experiments, face some common challenges to generalizing findings.

In those social sciences that have a methodological grounding in statistics, external validity is dependent on whether the inferences made from the results are based on a known probability (Bamberger, 2012; Biggeri and Ciani, 2012; Harding and Woolcock, 2012). This is typically achieved by using random, representative samples, and random assignment of treatments, exposures, or interventions, which thereby establishes the statistical basis for generalizability. Qualitative researchers who use less statistically-based methods are also concerned with the external relevance of their findings, and with which known conditions affect the external relevance (Dellinger and Leech, 2007; Groleau et al., 2009; Whittemore et al., 2001). Without a probability distribution as an analytical reference, qualitative researchers primarily address generalizability issues through the consistent application of thematic, contextual and conceptual frameworks that guide the transferability of findings, thereby ensuring the analytical basis for generalizability (Bamberger, 2012; Byrne, 2013; Yin, 2009, 2011).

Proponents of case study research and practitioners of qualitative approaches counter that there are other ways to think about generalizing findings than statistically, and corresponding systematic methods other than randomization and probability based sampling, to maximize external validity. These methods emphasize the theoretical foundations for conceptualizing the external relevance of research findings, or the analytical generalizability (Yin, 2009) and the transferability of findings (Burchett et al., 2011; Green and Glasgow, 2006), over the mathematical calculations that underlie statistical generalizability (Gerring, 2007; Seawright and Gerring, 2008; Yin, 2009). The objective

for qualitative research, therefore, is analytical generalizability, or ensuring confidence that findings are relevant to a similar situation, rather than statistical generalizability, where findings are considered relevant for the population from which the representative sample was drawn (Yin, 2009, 2011).

Understanding is often lacking across methodological disciplines about the common conceptual foundations of 'external validity' and 'transferability', and the rigour of approaches used to maximize each. However, when faced with interpreting findings from evaluations of complex interventions, a common approach emerges from both the quantitative and qualitative disciplines: using program theory, or theory of change, as the spine of the research approach (Lemire et al., 2012; Sridharan and Nakaima, 2012). Theory-based methodology also emerges as the preferred primary approach to maximizing generalizability of case study findings (Douthwaite et al., 2003; Judge and Bauld, 2001; Rogers, 2008; Rogers et al., 2004). A key question facing case study researchers is: what role can theory play in maximizing the generalizability of case study results?

This article considers how the Africa Routine Immunization System Essentials (ARISE) project addressed the problem of generalizing findings from a series of country case studies to understand recent improvements in immunization coverage in Africa, and what lessons were learned along the way. Specifically we focus on how program theory provided the foundation for systematic case selection and case analysis and synthesis processes, and what was achieved through theory, case selection, and case analysis and synthesis, with respect to maximizing the generalizability of the case study findings.

Background on the ARISE project

The ARISE project was created to learn from the countries whose immunization systems are performing well by documenting their experiences and consolidating them into a body of evidence that can inform future decisions on how to diffuse and scale up innovation. Established in late 2009, ARISE was managed by JSI Research & Training Institute, Inc. (JSI) and funded by the Bill & Melinda Gates Foundation.

It was known that African countries have achieved solid advances in immunization programming since 2000, with third dose diphtheria-tetanus-pertussis (DTP3) coverage in Africa increasing from 55 percent to 71 percent ('WHO/UNICEF estimates', 2010). In general, progress has been steady and sustained, but the underlying reasons for these improvements in immunization performance were not well understood. The first analytical initiative of ARISE (LaFond et al., 2011) identified that available data did not increase the understanding of what initiatives and processes were actually 'driving' the improvements in routine immunization. There was insufficient detail on how particular drivers were introduced, how they were implemented, and on the pathways by which they contributed to routine immunization performance improvement. In-depth country-level case studies were identified as early as 2006 (Naimoli et al., 2008) as a necessary step to better understand the links between program implementation and immunization performance.

The overarching research question for the ARISE country case studies was, 'What are the drivers of routine immunization system performance?' Specifically, the country case studies aimed to answer:

1. Which specific drivers affect the routine immunization system, and in what ways?
2. How are they implemented?
3. Why were these drivers successful in improving routine immunization performance?
4. How do specific contextual factors influence the effectiveness of routine immunization system performance drivers?

5. Who are the key stakeholders associated with the drivers?
6. Which drivers present viable investment options and by what possible mechanisms?
7. How can investments promote or establish drivers in other similar contexts?

For the purposes of the ARISE project, a routine immunization performance driver was defined as a structure, resource, or process that works on or through the immunization system to improve performance or perform effectively (LaFond et al., 2011, 2012) The focus of the investigation was on the process of performance improvement (as measured by recent changes in DPT3/Penta 3 coverage), rather than sustained high coverage. The intention was to generate detailed information on how identified drivers function in practice, the pathways through which the drivers influence performance, the contextual factors important to the drivers, and the relationships among different drivers identified. Because the ultimate objective was to provide direction for future investments in immunization, it was imperative that the case studies contribute generalizable findings.

Developing a theory-based analytical framework

As the theoretical foundation for the study approach, we developed a theory of change for the ARISE case studies (Figure 1), based on immunization program theory. We began with a basic logic model of the routine immunization system and then hypothesized how performance improvement drivers might influence the pathway from inputs to outcomes (in our case, coverage improvement). We expanded the model by positing that performance drivers could intervene at various and possibly multiple points along the pathway. We then included context as a potential influential factor along the entire pathway to performance improvement, defining 'context' as those factors that are outside the direct influence of the routine immunization system − including broader health system factors. This theory development process included the perspectives and input from a large team of ARISE stakeholders, which included immunization experts.

The theory of change and associated hypotheses provided the basis for case selection, data collection, and analysis. Because our study sought to understand how and why immunization coverage improved and to identify the drivers of performance improvement, the case selection strategy used the program theory's outcome, immunization system performance as measured by DTP3 coverage, as the first case selection criterion. Specific data collection exercises (the RI situation analysis, in-depth interviews and observation) documented aspects of the preceding components of the logic model – inputs, essential systems components, and processes – to be able to identify the performance drivers and determine where the drivers were intervening to improve performance. Data collection also sought to establish a deep understanding of the context in which the immunization program was being implemented. Within-district analysis was employed to categorize findings according to possible intervention points, and identify any ordering, preconditions, and interactions among the elements of the program theory. Cross-district analysis was employed to look for commonalities among the drivers of immunization performance improvement, using the program theory's hypothesized intervention points, ordering, and contextual factors to frame and limit the extent of interpreting a driver as 'common'.

Before case selection began, we recognized the challenge of verifying a driver's contribution to immunization coverage improvement. We also recognized that the limited time and budget available for data collection would prevent us from getting a deep enough understanding of the implementation context. Therefore, to validate findings and to gather data from a counterfactual situation, we also included negative cases, where coverage levels remained steady over the same time period.

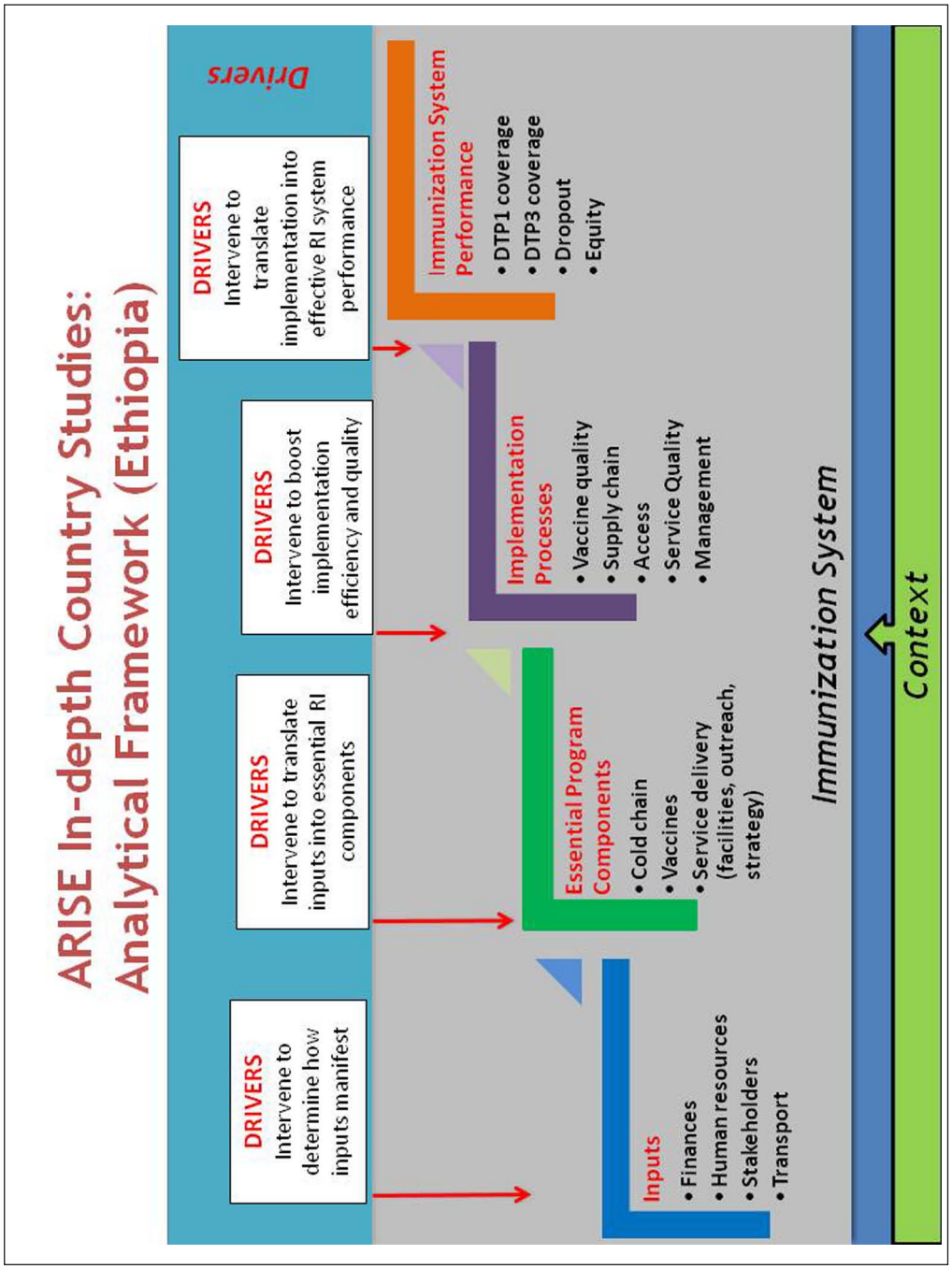

Figure 1. Analytical framework for ARISE case studies.
RI: Rhode Island

Case selection

Case selection was designed as a two stage process: first, countries were selected by the ARISE study team using a set of standard criteria, and then districts were selected as embedded cases within countries according to a selection protocol applied by the country study teams. In both steps, we used objective data first to define immunization performance, and then pragmatic and other considerations of accessibility, logistics, and safety.

Country case selection process

In order to understand what was driving improvements in coverage, ARISE chose to conduct research in countries where national DPT3 coverage had improved in the past 10 years (2000−09). In addition, our intent was to identify countries that were broadly representative in terms of the main characteristics of countries in sub-Saharan Africa. We first listed countries in sub-Saharan Africa based on reported DPT3 coverage from 2000−09. We then excluded countries where DPT3 coverage had not improved in recent years, or baseline coverage was too low (under 60%). Post-conflict countries were also excluded as not typical of the sub-Saharan African RI implementation context,[1] as were countries that were considered to be on the extremes of population and geographic size (again, because they were thought not to be typical of SSA). In this first step, we excluded 38 countries based on one or more criteria and reached a short list of 10 (ARISE Country Selection for In-depth Studies: A Guide to Country Selection, JSI, October 2010). After further consultation with stakeholders, three countries were selected for inclusion in the ARISE in-depth study: Cameroon, Ethiopia, and Ghana. At this point we also defined the unit of analysis as the district, rather than the country, because the district serves as the main administrative unit for immunization and other health programs. We determined that country contextual factors (immunization program history, financing, policies) would be used to interpret the implementation context for driver identification at the district level, and when synthesizing findings across the cases.

Embedded case selection process

In the second stage we selected three districts per country with recent improvement in immunization coverage from 2007−10 and one district where coverage had remained steady during the same period. Districts were eligible for selection as a 'positive change' district if they had at least 60−70 percent coverage three years prior (2006−07), and had reached over 80 percent DTP3 coverage by 2009−10. They were eligible as a 'steady' district if they had at least 60−70 percent coverage in 2006−07 and had experience little or no change in coverage over the same period. Focusing on recent program experience allowed us to reduce recall bias, and improve our chances of speaking with respondents who had been present during the period. To find cases that were typical of most districts and minimize other major sources of variation we purposively excluded all cases with unusually difficult geography; below average population density; and where there was a large urban center. In all but one country (Cameroon), we selected each of the positive change districts from a different region of the country to represent a range of regional contexts. We selected the negative case from one of the same regions as a positive case to minimize the effect of context on the negative case outcome.

Overall, we studied the experience of four districts in each of the three countries, for a total of 12 cases – nine positive and three negative.

We anticipated that this selection strategy would allow for easier comparison between the positive and negative cases, within a consistent set of contextual factors, to determine whether

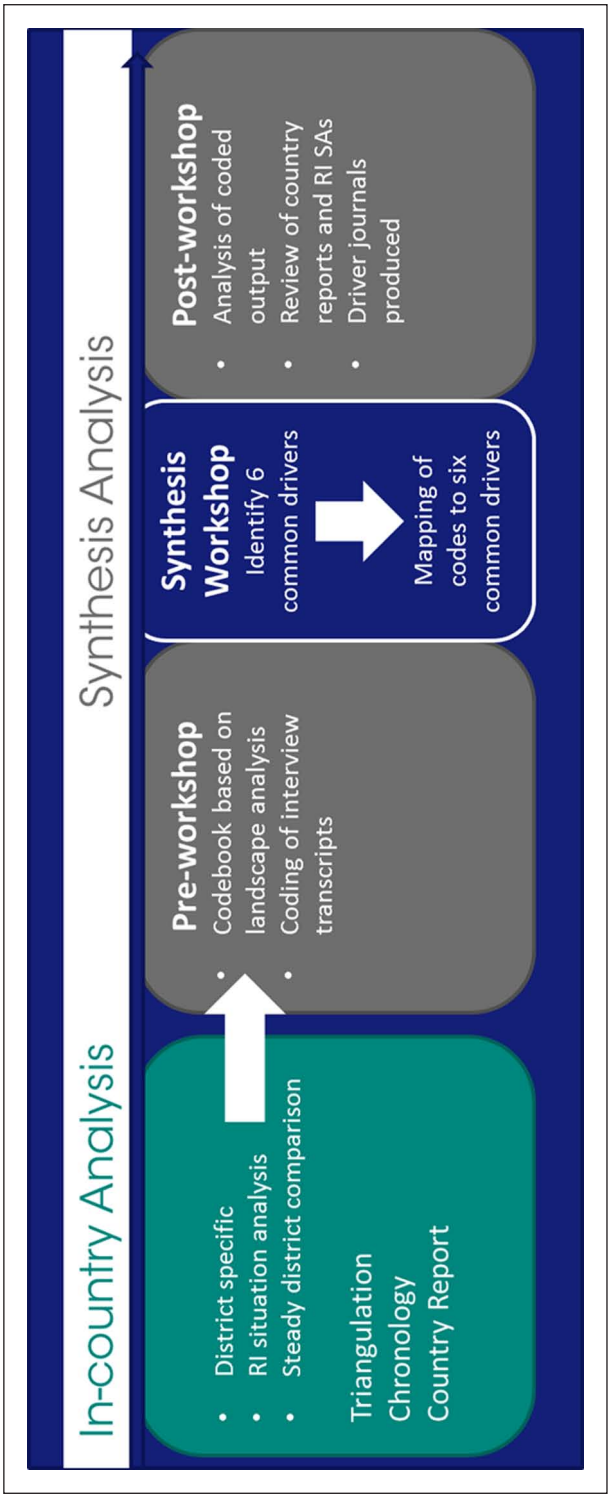

Figure 2. Overview of ARISE in-depth study analyses.
RI: Rhode Island

identified 'drivers' from the positive cases were present or absent in the negative cases. In all three countries, the selected districts were reviewed by immunization specialists to ensure that the apparent trends in coverage accurately reflected the performance of the districts' RI systems. In Cameroon, as fieldwork progressed, it became apparent that the district that had been selected as 'steady', had been performing very well for many years, contrary to what the data showed. This district was reclassified as a high performing district, and a new district with confirmed steady Penta3 coverage was added as the negative case.

Each country team took this selection criteria framework and implemented a case selection process in country. The country teams were challenged by case selection and the desire to maximize the generalizability of the case study findings to sub-Saharan Africa. They considered questions such as: What constitutes a 'typical district' in terms of immunization program performance? How does one categorize performance improvement experience? Is a district whose coverage improved from 40–60 percent the same as a district whose coverage improved from 60–80 percent? Should more 'typical' districts, be preferred, or those with greater performance improvement? Early on, the ARISE team engaged in continued debate, with internal and external EPI experts, about what cases would give the best chance of answering the research question, What drives immunization performance improvement, once the EPI program is properly resourced and functioning? EPI experts were able to determine that a starting performance level of less than 50 percent would likely indicate that the program was in need to strengthening, whereas a starting performance level of closer to 60 percent likely indicated that the program was well functioning. The EPI advisors to ARISE were also able to determine that, over a three-year period, a coverage increase of 15−20 percent was reasonable to expect, whereas anything more than that likely indicated something suspicious with the data or unusual investments or events. These refinements in defining a 'positive change' case were incorporated into the ToC and the case selection criteria, as well as later into the constructing the driver pathways to performance improvement, and in generalizing the synthesized findings.

Case analysis and synthesis

The objective of the ARISE in-depth case studies was to generate evidence to support recommendations for future investment that would be likely to improve immunization performance in sub-Saharan African countries. To that end, the findings needed to be synthesized across the individual cases. The analytical strategy was to first conduct within-case analysis of each district, then identify common themes and drivers based on the within-case findings from four districts in a country, and then pool the data across all 12 districts and conduct a cross-case synthesis analysis. The analysis process had three stages:

1. In-country analysis;
2. Cross-case common theme identification;
3. Cross-case synthesis.

In-country analysis

Each country case study team was responsible for the country level analysis, and identification of drivers and other country-specific findings. This was intended to proceed by first analyzing each embedded 'typical' case (the improving performance districts) for identification of important drivers of improved immunization performance, key contextual factors, and any other relevant themes.

The negative 'deviant' case (the steady performance district) was to be the last source of data, and used to validate whether the identified drivers were present, absent, or not functioning, why the drivers had that status in the negative embedded case, and what contextual factors might be influencing that status. Therefore the embedded cases were not intended to be treated as 'observations' in a data set, but each was treated as its own study, first analyzed for its particular conditions of variation and complexity.

In each country, district-by-district analysis was done, and theories and hypotheses developed, tested with data, and use to evolve subsequent data collection and analysis. How this transpired in each country varied, but each produced pathways of how the drivers contributed to improved immunization performance, chronologies of the drivers and their contributions to RI performance, or models of change used by the teams. All country teams strived to replicate findings on the drivers in the 3 positive districts, and conduct comparisons with steady district, including corroboration with program data and RI situation analysis findings. Each country team identified that replicating fieldwork in three districts whose coverage improved and in a fourth district without positive change allowed the researchers to test rigorously their hypotheses about the drivers of immunization performance change. Each team also shared preliminary conclusions with country stakeholders, received their feedback, and produced a stand-alone country case study report.[2]

In reality, the country case study teams often found it difficult, primarily because of a short time frame, to complete district-specific analyses in the recommended sequence during the data collection process, and the full analysis was often completed only after data collection was completed in all the four districts. However, analysis was still done on a district-by-district basis, and theory building was done in stages, before drivers were finally identified. Then the embedded cases were pooled by the teams, after they were fully analyzed for individual context and variation. This ensured that the embedded cases could also serve as observations, while the integrity of the individual country cases was maintained.

Cross-case common theme identification

To being the process of 'generalizing' the country-specific findings, ARISE conducted a 'Synthesis Workshop' approximately a month after the last country team had finished their analysis and draft report. The goal of the workshop was to take findings from all three study countries, and determine and define what could be considered as 'common' drivers of improved routine immunization performance, while accounting for the individual variation in each case. The 21 workshop participants in the three-day workshop included members of each country case team, the ARISE synthesis team, and senior ARISE advisors.

The first day of the synthesis workshop focused on identifying common drivers across the 12 districts. First, all the drivers identified by the country teams were listed, mapped by district characteristics, and categorized according to themes. Discussion about the 'essence' of the drivers was encouraged – participants were asked to think about, if you stripped away the implementation variations, what was the driver really about? From these discussions, the analytical foundation for six common drivers emerged (Table 1). For example, the Ethiopia study team had identified 'Accountability and performance monitoring' as a critical driver of the improvements in immunization in their three positive change districts, and had validated that this had not taken hold yet in the negative case district. Through further discussion of the essence of this driver, it became clear that the key element of accountability and performance monitoring in Ethiopia was equivalent to the regular review meetings that the Ghana team identified as a driver. Further, 'regular review meetings' encompassed the combination of 'management systems' and 'HW motivation' drivers

Table 1. Six common drivers of RI performance improvement, linked with country-specific drivers.

Common driver	Ghana drivers	Ethiopia drivers	Cameroon drivers
1. Political and social commitment to routine immunization	Essential EPI infrastructure Community engagement	Active community participation	Community Involvement – stakeholders; Systems Foundation – strategic approach
2. Actions of development partners	Essential EPI infrastructure	Support from development partners	Systems Foundation – strategic approach
3. Cadre of community-centered health workers	Essential EPI infrastructure	Locally recruited and supported HEWs	System Foundation – implementation resources
4. Health system and community partnership	Community engagement	Partnership between DHMT and local government	Community Involvement − stakeholders
5. Regular review of program and health worker performance	Clear roles and responsibilities Regular review meetings	Accountabilty and performance monitoring	Systems Foundation – management systems; Human Factor Catalysts – HW motivation
6. Immunization services tailored to community needs	Autonomous local health teams Community engagement	Active community participation Partnership between DHMT and local government	Service Delivery – strategy, quality, infrstructure; Human Factor Catalysts – HW capacity

that were identified by the Cameroon team. Therefore, the common driver of 'Regular review of program and health worker performance' was defined in a way that accurately reflected the essence of these drivers in all nine districts and the driver, as defined, was considered 'common'.

Six drivers of RI performance improvement emerged as common to the nine study districts where immunization coverage improved (Table 1):

1. Political and social commitment to routine immunization;
2. Actions of development partners;
3. Cadre of community-centered health workers;
4. Health system and community partnership;
5. Regular review of program and health worker performance;
6. Immunization services tailored to community needs.

Although the exact way in which these drivers contributed to improved coverage varied somewhat by district, each was present in the nine positive change districts and was either absent or weaker in the three study districts where coverage remained steady. In addition to identifying, exploring, and categorizing the common drivers of improved coverage in each district, we wanted to understand how the district implementation context influenced both driver effectiveness and coverage improvement, and which elements of the health and immunization system facilitated the work of the identified immunization performance improvement drivers.

The second day therefore focused on identifying contextual factors and how they interacted with the drivers. Workshop participants worked in small groups that had representatives from each country team, each working with a specific common driver, to first identify and define the full range of contextual factors that were found to affect the driver. These contextual factors were

categorized as: essential for the driver to work; facilitates the driver, but not necessary. Then the time relationship to the driver was determined – specifically, was the contextual factor a necessary pre-condition for the driver or not. First, common to all districts where coverage improved was the solid foundation provided by the regular availability of the essential components of the immunization system: trained staff, vaccines and cold chain, transportation, and sufficient sites from which to deliver immunizations. Without this foundation, many of the driver mechanisms could not perform effectively and facilitate coverage improvement. Second, in districts where coverage improved, managers and health workers focused delivery strategies on reaching the community. Third, in positive change districts, the teams tailored these delivery strategies to local conditions and needs. Fourth, district experiences with coverage improvement suggested that the drivers do not operate in isolation but instead feed into and support one another in various ways.

Because each selected case actually had slightly different starting Penta3 coverage levels, we also analyzed whether the driver seemed to be more or less 'present', or 'important', by the different levels of starting point Penta3 coverage (50−60%, 60−80%, or above 80%). Finally, the small groups determined what levels of the RI system were involved with the driver, and with the associated contextual factors: national, regional/zonal, district, facility, community, global. These results were integrated with the common drivers listing and description from the previous day.

On the third day, chronologies of the pathways for the six common drivers were developed to test that these maintained logical sense when synthesized across the nine districts and contextual situations. Finally, the six common drivers were mapped back to the theory of change (Figure 3). The synthesis workshop essentially validated the original theory of change.

The results from the synthesis workshop, particularly the descriptions of the six common drivers, the analysis of contextual factors and the preliminary pathways from each driver to improved immunization performance, provided the analytical focus for the next stage of synthesis analysis, which pooled the all data from all 12 districts.

Cross-case synthesis

The data from each case was thematically coded using NVivo 9 software as a data management tool, which allows for the storage, coding and sorting of data. Because of changes in the timeframe for completion of the country cases study reports, the codes were originally derived from the results of the Landscape Analysis, background documentation and expert opinion. Synthesis analysis began by mapping the previously derived NVivo codes to the six common drivers identified during the workshop. This mapping exercise was used to inform decisions about which codes to include when querying for data related to each of the six common drivers using Nvivo.

Once the coded data relevant to each of the six common drivers was generated, the data was organized using an analysis format termed 'Driver Journal'. Typically, each analyst read through the data a minimum of four times, each time with a slightly different analytical objective. The first objective was to become fully familiar with the particulars of the coded data, and how much data there was in support of the common driver description. Themes regarding the essential components of the driver were preliminarily documented at this stage. The second read-through focused on the objective of finding the data that described, in detail, the pathway through which the driver exerted its positive influence on immunization coverage. The third read-through focused on the data from the steady districts, to identify data supporting the absence or presence of the driver, and if it was present, data on what made that driver different or not as effective in the steady district. The fourth read through focused on finding specific quotes that best illustrated the essence of the driver and its pathway to performance.

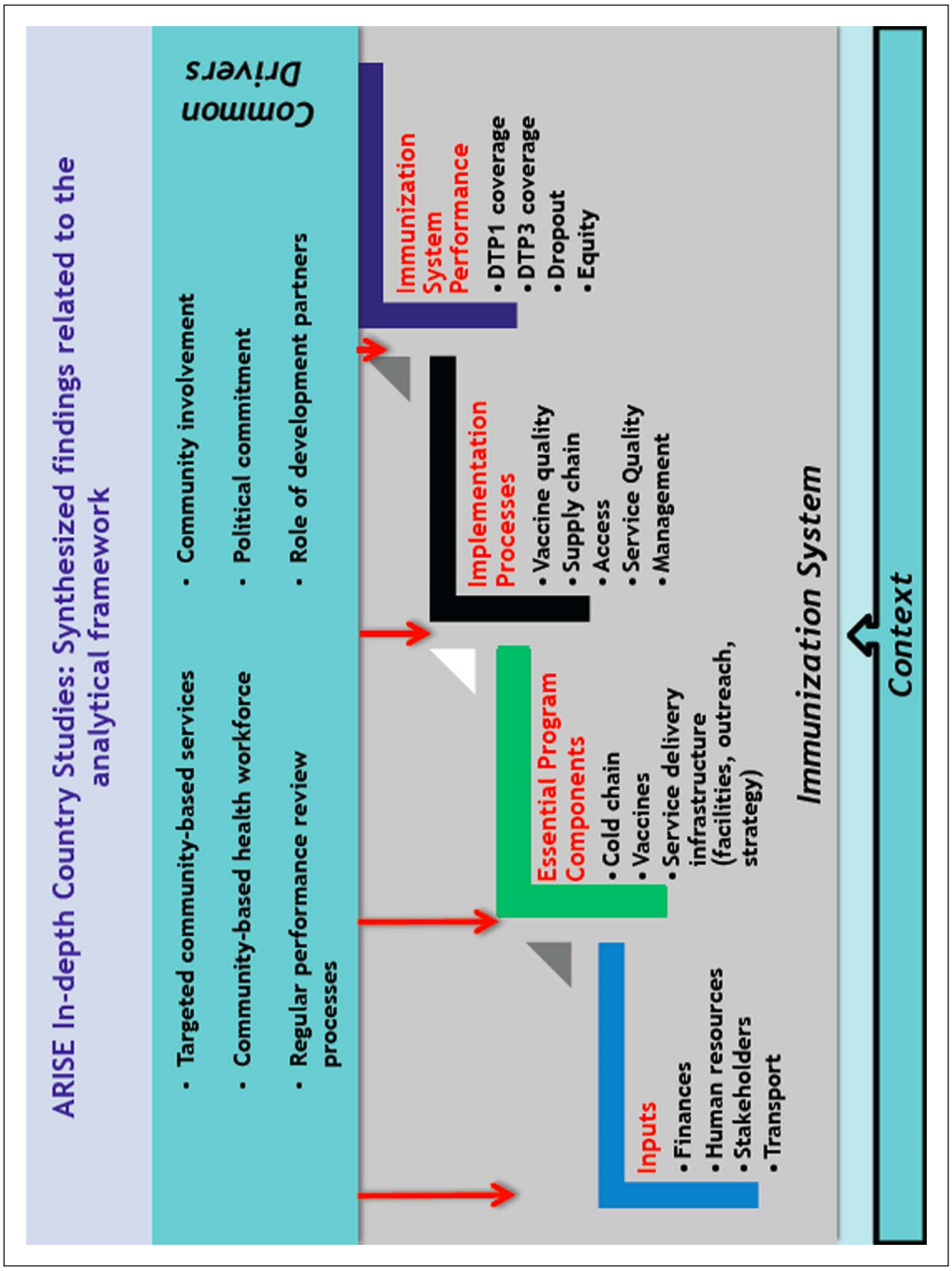

Figure 3. Common drivers related to the ARISE analytical framework.

The synthesis analysis team also reviewed the individual country case reports, national RI situation assessments, the detailed notes from the Synthesis Workshop, and the chronologies produced by both the country team and during the Synthesis Workshop, and integrated the relevant information into the Driver Journal analysis format. In putting all of this together into a driver journal for each of the six common drivers, the synthesis analysis team endeavored to lay an evidence 'trail' that would maximize the transparency of the synthesis analysis process and the resulting conclusions regarding the relationships among the drivers, their essential components, and their pathways to improving immunization performance.[3]

Because the synthesis workshop had pointed to the possibility that the drivers did not function in isolation, but had synergies and interdependencies, the synthesis analysis team also worked intensively to outline how the drivers were inter-related, and to review the coded data again to determine whether the inter-relationships could be considered clustering.

The analysis of the pooled data showed that four key aspects of the district implementation context were found to support the improved performance of the RI system: 1) decentralization of the health system to at least the district level (including the establishment of district health teams to manage decentralized health services); 2) a recent policy or strategy that extended primary health care services through a community-based network; 3) the introduction or implementation of an integrated package of basic health services which includes RI; and 4) the prioritization of DTP3/Pentavalent3 coverage as a key indicator of health and development.[4] These contextual variables were present to some degree in all 12 study districts, both those with recently improved coverage and those with steady coverage.

In this process, the team also was able to determine that two drivers − political and social commitment to RI and the actions of development partners − are more accurately considered *enablers* of RI coverage improvement, rather than as forces that directly influence performance change.

Figure 4 shows the final graphical representation of the identified synthesized drivers of RI performance improvements, including the ordering and clustering. The two enablers, political commitment and actions of development partners, precede the two pre-conditions of basic RI resources and capacity and district management teams being in place. In parsing the implementation context across the 12 cases, it was clear that no immunization performance improvement could occur without these two pre-conditions. Further, sufficient political and social commitment, along with supporting action from development partners, were needed to establish these pre-conditions – otherwise investments in RI would not be sufficient to support any performance improvements. The four direct drivers of RI performance improvement cluster together – the synthesized analysis showed that, when a cadre of community-based workers are present, and their performance is sufficiently supervised and reviewed, they often forge a partnership with the community to design and deliver immunization services specifically tailored to that community's needs – and this is the most direct pathway to increased acceptance of services and better delivery of immunizations, which results in improved coverage.

Generalizability of ARISE case study findings – how far could we go?

ARISE's objective was to generate findings that could be used to improve immunization performance across Sub-Saharan Africa, by better understanding what had driven performance improvements in some areas of SSA, and under what circumstances. To do this, ARISE had aimed for regional variety in country contexts (East, West, and Francophone Africa), omitting the special situation of the post-conflict country, and then selected cases within these country contexts such that they were more likely to be similar to the typical RI implementation situation at district level.

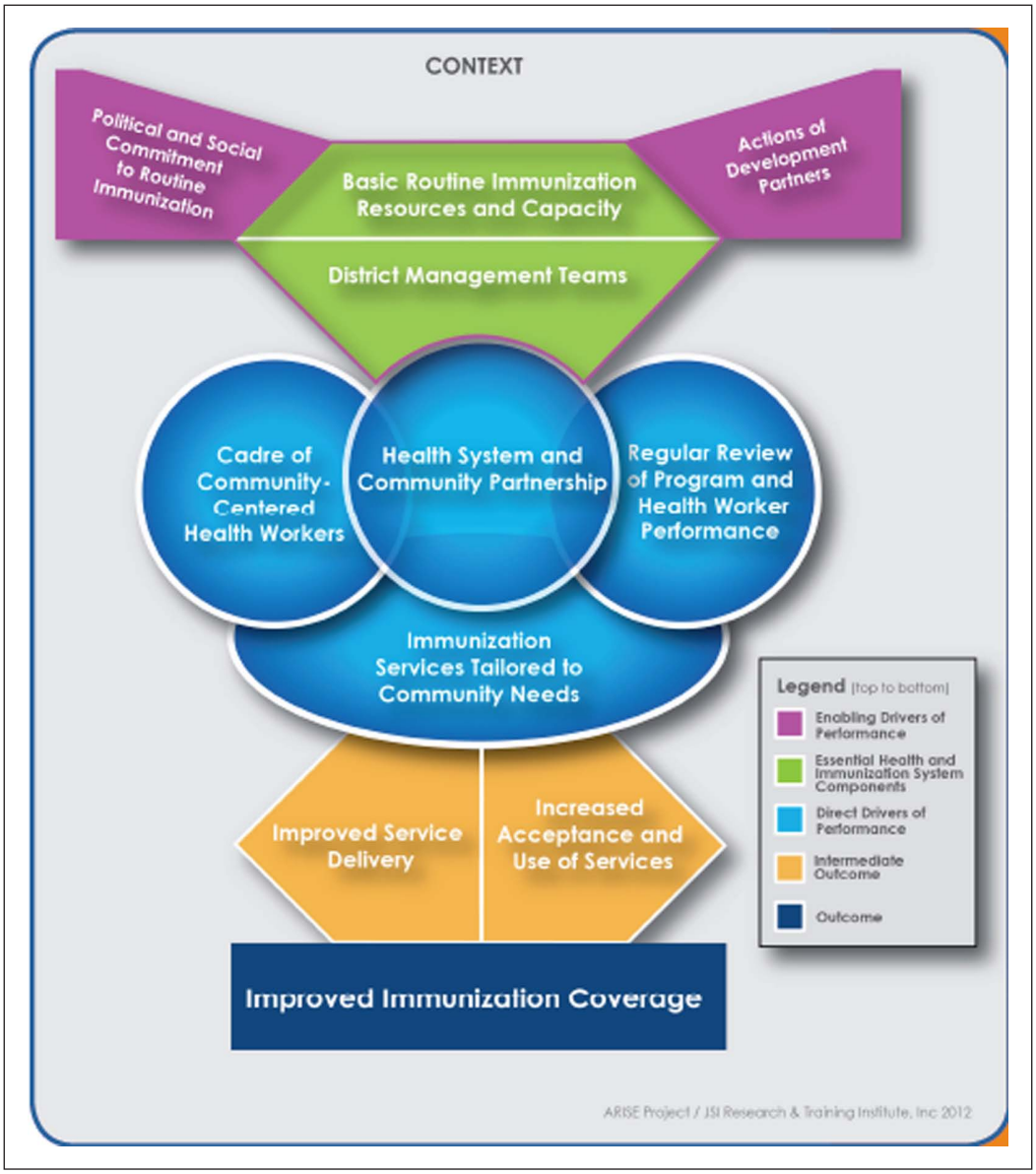

Figure 4. Final synthesized findings on drivers of improved RI performance, ARISE, showing clustering and ordering.

This made analytical sense, as it is by and large the districts in SSA that manage health service delivery, and are, more and more, held responsible for health systems performance.

The theory of change was the foundation for the study design, data collection, and different stages of analysis, all using the district as the unit of analysis. As we were interested in cases where RI performance had improved, our baseline coverage rate for eligibility was a minimum of 60 percent. Our synthesis analysis showed us that the first two steps in the theory of change, inputs

and essential program components, were important pre-conditions for any expected RI performance improvement. Further, our final analysis identified another important pre-condition, the existence of functioning district health management teams. Two enabling factors, political and social commitment and actions from development partners, were found to be necessary for the pre-conditions to be met in any implementation context. These case selection criteria, implementation pre-conditions, contextual and enabling factors framed the four conditions for generalizing the ARISE findings: 1) baseline DPT3/Penta3 coverage rates of 60 percent or more; 2) political and social commitment to RI; 3) basic EPI resources and capacity are in place; 4) district health management teams are in place. Immunization experts on the ARISE team, and on the External Experts Advisory Group to ARISE, felt that, given these conditions, the results were relevant to most districts in sub-Saharan African countries.

The generalizability of the ARISE findings can be clearly articulated as a description of Figure 4. In contexts where: 1) the baseline DTP3/Penta3 coverage is 60 percent or more; 2) there is sufficient political and social commitment and development partner action to ensure that 3) the basic EPI resources and capacity have been established; and, 4) the country has district level management teams in place, then ensuring a cadre of community-centred health workers is in place, ensuring their RI performance is sufficiently supported, supervised, and reviewed; and, facilitating partnership with the community to ensure that appropriately targeted RI services are delivered will increase acceptance of RI in the community, improve the delivery of RI services, and ultimately, result in increased RI performance improvement in terms of coverage.

Methodological lessons learned

What contributed to the generalizability of the findings from the ARISE in-depth case studies? First, a strong theory of change that was the foundation for the study design, data collection, and the within-case and cross-case analyses provided the conceptual touchstone for each step of data collection and analysis. It was critical for establishing consistency in data collection and analysis approaches used by three different case study teams, which then allowed for cross-case comparability.

Second, clear case selection criteria, systematically applied, strengthened our ability to generalize findings beyond the cases. The theory of change helped identify the first case selection criteria, baseline RI coverage and improvement over three years. The need to generalize to the 'typical' implementation situation in SSA gave focus to the implementation context situation that should be considered as cases, and which contexts should not, further identifying selection criteria. These selection criteria became two of the four conditions for generalizing ARISE findings.

Third, the use of the negative case was a important feature of the study design that allowed for testing of theories and hypotheses regarding the drivers and their pathways to improving RI performance. The use of multiple cases, each with embedded typical cases and an embedded negative case, demonstrated sufficient replication of the findings for immunization stakeholders to accept the drivers by the ARISE case studies as a generalized phenomenon. The negative case was used for within-case analysis, for identifying the common themes across the cases, and during pooled analysis of the cases. By testing findings against a 'counterfactual' at many levels of analysis, confidence in the internal validity of the findings was strengthened, which provided stronger grounds for generalizing outside the cases.

Fourth, maintaining a balance between within and cross-case analysis was important, for ensuring internal validity of the findings. The implementation and contextual variation among the cases was not brushed under the rug, but treated and analyzed as part of interpreting findings

and defining the drivers that were ultimately identified as common. Sufficient replication was demonstrated by the way ARISE constructed the within-case analysis and cross-case synthesis – data were not pooled across cases, but each case was first analyzed for itself (Bennett and Elman, 2006; Byrne, 2012; Marchal et al., 2010; Yin, 2009). Findings from each case were examined to see if they were repeated or replicated in a workshop setting that included members from all the country case teams. Only after verification of this replication was data pooled for further analysis based on the replicated themes to produce the evidence base for synthesized findings that could be generalized to other sub-Saharan African countries. This built confidence in the synthesized findings, and provided the analytical foundation from which the synthesized findings could be generalized.

However, the case selection process proved to be quite difficult. Using immunization coverage data at the district level proved to be quite problematic at times, and the district selection protocol often had to be adapted in real time by the country teams. At times, selection criteria were found to be too restrictive (e.g., the coverage range was not wide enough); other times, not enough districts met the criteria. In some instances, districts originally selected as typical 'improving' cases were found to be steady, negative cases, or vice versa. When this happened, the country teams had to repeat the selection process, and expended valuable time in doing so. In most situations, the performance data that was the main case selection criterion could only be verified once the team had been in the districts and examined the coverage data at a very disaggregated level.

This did not invalidate the research approach used, but it showed that even when purportedly reliable program data are available, such as data routinely reported and tracked internationally for immunization performance, using these data for case selection does not necessarily guarantee a 'clean' process. Lacks of *a priori* data on context and implementation situations also meant that ARISE could not use these purposively for case selection, which would have contributed greatly to the analytical generalizability objectives.

Without a systematic case selection process that uses transparent case selection criteria derived from an analytical framework, most case studies use practical factors to determine case selection, such as time, money, access, and perhaps also the prominence of a particular case (Seawright and Gerring, 2008). ARISE used, to the extent possible, explicit, objectively verifiable selection criteria and a systematic process to establish the methodological justification for why a case is included, and to lay the groundwork for generalizability of the study findings. Without this, researchers must instead rely on a retrospective assessment of how their selected cases relate to the wider population of cases, without which their results may be subject to significant bias (Seawright and Gerring, 2008). In fact, the experience of the ARISE project was that pragmatic considerations had to enter into case selection, although objective criteria were used first, and stakeholder judgment and on-the-ground data verification were done in order to adhere as closely as possible to the case selection protocol; this was considered critically important for the ultimate generalizability objectives of the ARISE case studies. These challenges to ARISE's systematic case selection process suggest that case selection may need to be a stand-alone, formative part of case study research.

The theory of change developed prior to the start of data collection originally was intended to organize the synthesis analysis to categorize findings according to possible intervention points, and identify any ordering, preconditions, and interactions among the elements of the analytical framework. In reality, the drivers that were found to be replicated among the cases were sufficiently broad as to intervene along multiple points in the adapted logic model (Figure 3). Where the analytical framework proved most useful for analysis was in assessing the contextual factors, inputs, program components, and process situation in each case (i.e., for systematically assessing the context and implementation issues that influenced the transferability, or generalizability, of the findings).

Analyzing context also proved to be challenging. Despite the inclusion of key definitions of contextual factors in the theory of change, and the focused analysis during the Synthesis Workshop, we actually were limited in our ability to conduct in-depth analysis of all key contextual factors because of the sheer number and complexity of relationships and timing. Instead, we were only able to draw out four common contextual factors that were conducive to immunization performance improvement, not because these were completely absent in the steady district, but because the data showed that these were named in the positive change districts as potential drivers or enablers of drivers. We also focused on these four factors because they had figured prominently in our ToC development, and the assumptions and hypotheses underlying the ToC. We concentrated our analysis on establishing confidence in the commonalities of findings as the path to generalizability, and the role of the four identified principal contextual factors in affecting that generalizability path. A more complete analysis would require more time and possibly follow-up data collection to be able to map accurately the relationships among all the contextual factors themselves, and their influence on the identified drivers of immunization performance improvement.

The use of negative embedded cases in the ARISE case study design was greatly appealing to stakeholders, especially those most familiar with quasi-experimental research and evaluation designs. Although the negative embedded cases were critical to the study design, the country teams sometimes found them difficult to use, especially for presenting evidence from the negative cases, as the 'evidence' was often simply the absence of evidence. The teams needed more methodological guidance and data collection and analysis tools to be able to more fully realize the methodological potential from the negative embedded cases. It was relatively more straightforward to use the pooled data from the negative cases in the synthesis analysis to determine whether the six common drivers were absent, present, weak, or changed in form. Regardless, in reviewing the literature, it appears that there is a general need for more documentation on how negative cases have been used in case study that could contribute to practical guidance.

Although the multiple case study design, with embedded typical and negative cases, was a powerful approach to maximizing the generalizability of the findings, and importantly, convincing stakeholders of the generalizability of the case study findings, integrating the complexity found within each case into findings that are synthesized across cases was challenging. The critical step was to include implementation nuances and context – including history – in the interpretation of findings. In the ARISE project, this was done through a series of participatory analysis exercises in a transparent workshop setting that included members from all the country case teams that contributed to further confidence in the generalizability of the synthesized case study findings. The use of repeated case studies by the ARISE project therefore supported the development of theories that incorporated specifics on immunization program implementation and context situations, increasing the generalizability of the case study findings, and improving the understanding of the causal pathways between the identified drivers and measured immunization performance.

Conclusions

One of the tentative conclusions from the International Workshop on Case Studies in Development Evaluation was that there could be two main ways to draw conclusions and propose causal pathways based on a relatively small number of cases ('Case Studies in Development Evaluation: Validity, Generalization and Learning. An International Workshop', 2012). The first was to place emphasis on exploration of internal relationships within a case, as well as comparisons across cases, in order to address the key issue of case complexity as a precursor to generalizing findings.

The second was to give theory a central role, whether for testing programs against theory or to develop theory as part of a case study evaluation.

ARISE did both of these things. Theory played a central role in the ARISE case studies. The theory of change provided ARISE with the ability to address the complexity of the context and implementation processes of each individual case explicitly, which aided immeasurably with identifying replication of findings across cases and provided a solid methodological grounding for the synthesis. In turn this greatly improved the confidence in the generalizability of the findings. However, the ARISE team still had to endeavor to find effective ways to communicate the importance of individual case complexity when interpreting the findings. The need to refer to the underlying theory of the ARISE case studies added a layer of complexity when communicating findings and their implications, which is likely to be a burden for most theory-driven evaluations, whether case study-based or not. The additional need to use deductive logic, based on expectations about the causal relationships of interest and the characteristics of the selected cases, to lend explanatory weight to the case study analysis, added another layer of complexity for consumers of the ARISE case studies to understand how generalizable the findings may be.

The experience of the ARISE project suggests that a balance between within-case and cross-case analyses is necessary to adequately accommodate internal relationships found in individual cases to maximize the generalizability of the findings. This points to a recommendation that case studies that use multiple cases and intend to synthesize findings should sequence the analysis process to allow:

- first for within-case analysis;
- then identification of replicated findings across cases which includes consideration of context and implementation variations; and
- finally synthesis across cases.

Without this sequencing, the risk is that relationships internal to an individual case are overlooked in synthesis, and the context and the complexity of each case are lost or diminished, and the generalizability of the findings becomes questionable.

Finally, the ARISE case study experience is that the case selection process, including the appropriate identification and application of selection criteria, functions as a critical methodological input to maximizing generalizability. Transparency and the use of objective criteria in the case selection process were central to establishing the foundations for generalizing findings. Our experience further showed that this aspect of case study methodology requires a dedicated process that should be planned separately, even when based on objectively verifiable quantitative measures, and should sufficiently precede data collection.

The International Workshop on Case Studies in Development Evaluation, as noted in the 'highlights' document, was an important opportunity to review the 'state-of-the-art' in how case studies are being used in development evaluation. It was clear that much thinking has been done on how to improve the quality and validity of case studies, and that interesting solutions to some common problems were beginning to emerge. The ARISE case studies propose that systematic, transparent case selection, and balanced within- and between-case analyses, both based on theory, are ways in which case study evaluation methodology can be strengthened.

Funding

This research received no specific grant from any funding agency in the public, commercial or not-for-profit sectors.

Notes

1. Post-conflict countries were excluded due to variability in their immunization coverage in the period post-conflict. Post-conflict countries are known to achieve/restore public health services through discrete phases including initial humanitarian aid followed by development aid. In post-conflict settings, immunization campaigns are often used to reach large numbers of children quickly including those who were unreached during the conflict. Therefore, immunization coverage in post-conflict settings may initially 'spike' and then return to low levels.
2. Available on http://arise.jsi.com/technical-resources/
3. The results on the drivers of immunization performance and proposed causal pathways are presented elsewhere: LaFond et al. (2012).
4. This indicator had gained prominence in national health plans, in the GAVI Alliance funding agreements, and generally in the global context of commitment to achieving the Millennium Development Goals.

References

Bamberger M (2012) Introduction to mixed methods in impact evaluation. Impact Evaluation Notes No.3, August. URL: http://www.interaction.org/sites/default/files/Mixed%20Methods%20in%20Impact%20Evaluation%20%28English%29.pdf

Bennett A and Elman C (2006) Qualitative research: recent developments in case study methods. *Annual Review of Political Science* 9: 455–76.

Biggeri M and Ciani F (2012) The impact of community-based programmes: learning from two case studies of community based rehabilitation. Available at: http://um.dk/en/danida-en/results/eval/eference-documents/workshp12/

Burchett H, Umoquit M and Dobrow M (2011) How do we know when research from one setting can be useful in another? A review of external validity, applicability and transferability frameworks. *Journal of Health Services Research & Policy* 16: 238–44.

Byrne D (2013) Evaluating complex social interventions in a complex world. *Evaluation* 19(3).

Case Studies in Development Evaluation: Validity, Generalization and Learning. An International Workshop (2012) URL (consulted February 2013): http://um.dk/en/danida-en/results/eval/reference-documents/workshp12/

Dellinger AB and Leech NL (2007) Toward a unified validation framework in mixed methods research. *Journal of Mixed Methods Research* 1: 309–32.

Douthwaite B, Kuby T, Van De Fliert E and Schulz S (2003) Impact pathway evaluation: an approach for achieving and attributing impact in complex systems. *Agricultural Systems* 78: 243–65.

Gerring J (2007) *Case Study Research: Principles and Practices*. Cambridge: Cambridge University Press.

Green LW and Glasgow RE (2006) Evaluating the relevance, generalization, and applicability of research issues in external validation and translation methodology. *Evaluation & the Health Professions* 29: 126–53.

Groleau D, Zelkowitz P and Cabral IE (2009) Enhancing generalizability: moving from an intimate to a political voice. *Qualitative Health Research* 19: 416–26.

Harding A and Woolcock M (2012) But how generalizable is that? A primer for development practitioners on assessing the external validity of 'complex' interventions. World Bank. URL: http://um.dk/en/~/media/UM/Danish-site/Documents/Danida/Resultater/Eval/Ref%20doc/Woolcock.ashx

Judge K and Bauld L (2001) Strong theory, flexible methods: evaluating complex community-based initiatives. *Critical Public Health* 11: 19–38.

LaFond A, Kanagat N, Sequiera J, Steinglass R, Fields R and Mookherji S (2012) Drivers of routine immunization system performance at the district level: study findings from three countries. Research Brief No. 3. Arlington, VA: JSI Research & Training Institute, Inc., ARISE Project for the Bill & Melinda Gates Foundation.

LaFond A, Sequiera J, Kiwanuka S, Pomeroy A, Posner J and Macabasco R (2011) ARISE landscape analysis synopsis: an initial investigation of the drivers of routine immunization system performance in Africa. JSI. URL: http://arise.jsi.com/files/2012/11/Arise_LASynopsis_final508.pdf

Lemire ST, Nielsen SB and Dybdal L (2012) Making contribution analysis work: a practical framework for handling influencing factors and alternative explanations. *Evaluation* 18: 294–309.

Marchal B, Dedzo MD and Kegels G (2010) A realist evaluation of the management of a well-performing regional hospital in Ghana. *BMC Health Services Research* 10: 24.

Naimoli JF, Challa S, Schneidman M and Kostermans K (2008) Toward a grounded theory of why some immunization programmes in sub-Saharan Africa are more successful than others: a descriptive and exploratory assessment in six countries. *Health Policy and Planning* 23: 379–89.

Rogers PJ (2008) Using programme theory to evaluate complicated and complex aspects of interventions. *Evaluation* 14: 29–48.

Rogers PJ, Petrosino A, Huebner TA and Hacsi TA (2004) Program theory evaluation: practice, promise, and problems. *New Directions for Evaluation* 2000: 5–13.

Seawright J and Gerring J (2008) Case selection techniques in case study research a menu of qualitative and quantitative options. *Political Research Quarterly* 61: 294–308.

Sridharan S and Nakaima A (2012) Towards an evidence base of theory-driven evaluations: some questions for proponents of theory-driven evaluation. *Evaluation* 18: 378–95.

White H and Phillips D (2012) Addressing attribution of cause and effect in small n impact evaluations: towards an integrated framework. International Initiative for Impact Evaluation. Working Paper 15. URL: http://www.3ieimpact.org/media/filer/2012/06/29/working_paper_15.pdf

Whittemore R, Chase SK and Mandle CL (2001) Validity in qualitative research. *Qualitative Health Research* 11: 522–37.

WHO/UNICEF estimates (2010). Available at: apps.who.int/immunization_monitoring/globalsummary/timeseries/tscoveragedtp3.html

Yin RK (2009) *Case Study Research: Design and Methods*, 4th edn. Thousand Oaks, CA: SAGE.

Yin RK (2011) *Applications of Case Study Research*. Thousand Oaks, CA: SAGE.

Sangeeta Mookherji (PhD, MHS) is Assistant Professor, Department of Global Health, George Washington University; Director, Program Design, Monitoring and Evaluation Concentration; Methodology Advisor, ARISE.

Anne LaFond (MSc CHDC) is Director of the JSI Center for Health Information, Monitoring and Evaluation and Principal Investigator, Africa Routine Immunization System Essentials (ARISE).