

ORIGINAL

Performance & Impact Evaluation (P&IE) Design Report



at the UNIVERSITY *of* CHICAGO

PRESENTED TO:

USAID/Uganda
Joseph Mwangi

PRESENTED BY:

NORC at the University of Chicago
Jeffrey Telgarsky
Senior Vice President, International
Projects
4350 East-West Highway, 8th Floor
Bethesda, MD 20814
Telephone: (301) 634-9413
Fax: (301) 634-9301

TABLE OF CONTENTS

A.	THE SCHOOL READING AND HEALTH PROGRAM (SHRP).....	3
B.	APPROACH TO PERFORMANCE AND IMPACT EVALUATION.....	4
C.	PERFORMANCE EVALUATION.....	5
	C.1 Approach	
	C.2 Data Sources	
D.	IMPACT EVALUATION.....	10
	D.1 Impact Evaluation Design	
	D.1.1 District-level Comparison Group	
	D.1.2 Assignment to Treatment and Control Groups and Measurement of Impact	
	D.1.3 Estimating Impact	
	D.2 Impact Indicators and Data Sources	
	D.2.1 Result 1 – Reading Program	
	D.2.2 Result 2 – School Health Program	
	D.3 Sampling	
	D.4 Data Collection Plan	
	D.4.1 Result 1 – Reading Program	
	D.4.2 Result 2 – School Health Program	
E.	DATA QUALITY ASSESSMENT.....	28

LIST OF ANNEXES

ANNEX 1: Sample size calculation

ANNEX 2: Data quality assessment checklist

PERFORMANCE AND IMPACT EVALUATION (P&IE) OF THE USAID/UGANDA SCHOOL HEALTH AND READING PROGRAM

NORC at the University of Chicago (NORC), in partnership with the Panagora Group, is pleased to submit this Evaluation Design Report to USAID/Uganda for the Performance and Impact Evaluation (P&IE) of the five-year School Health and Reading Program (SHRP) in Uganda.

A. THE SCHOOL READING AND HEALTH PROGRAM (SHRP)

RTI is implementing the School Health and Reading Program as two separate activities:

- 1) Result 1: Improved Early Grade Reading and Transition to English
- 2) Result 2: Improved HIV/AIDS Knowledge, Attitudes and Practice

Result 1: Improved Early Grade Reading and Transition to English

For the Result 1 intervention, SHRP will focus on the nexus of language, pedagogy, and instructional materials to significantly improve students' early grade reading and P3 literacy scores, as well as bring to scale a "Ugandan led 'reading policy'" (RTI International, 2012, p. 1). The Early Grade Reading (EGR) intervention will be implemented at multiple levels.

- At the school level, the intervention will provide training to teachers in early grade literacy instruction using students' mother-tongue in P1-P3 and with a transition to English in P4;
- At the district level, instructional and assessment materials will be developed for P1-P4 in the students' mother tongue;
- At the national level, MOES systems and pedagogical and language framework will be strengthened to support mother-tongue based EGR and transition to English. Support will also be provided to strengthen policies related to reading, as well as increase advocacy for reading at multiple levels (e.g. student, teacher, school, district, and national).

Together, these interventions are expected to improve the instruction and learning environment of students and eventually lead to improved literacy skills.

The Result 1 intervention will implement teacher trainings using the district education structure, through Coordinating Center Tutors (CCTs) who are school support workers in charge of monitoring education quality within their Coordinating Centers (CC). Each CCT is responsible for a certain number of schools within a district (one district typically has multiple CCTs). The CCTs selected for the intervention will receive training directly from RTI and in turn deliver teacher training and program support in their schools, thus following a Training of Trainers (TOT) model.

In Year 1 of the intervention, SHRP will be working in 4 local languages (Luganda, Runyankore/Rukiga, Ateso, and Leblano) in 11 districts. Throughout the life of the project, SHRP will be working in a total of 12 local languages (4 languages in the first year, 4 additional languages in the second year, and 4 additional languages in the third year). In Year 1, SHRP will work in 410 schools: approximately 20 CCTs and 1,200 teachers will be trained.

- a. At the school level, the main activities planned are teacher trainings provided through CCTs. These trainings will focus on pedagogy with an emphasis on using structured lesson plans and learner books. These lesson plans will provide teachers with a practical step-by-step process for implementing the transitional bilingual approach mandated by the Ugandan EGR policy. RTI plans to develop variations of the intervention in order to use experimental approaches to inform scale up.

SHRP will thus implement three slightly different interventions, or “treatment arms”, in the first year of the project, as follows¹:

- a. Basic Program: teacher trainings + materials + some CCT support
- b. Basic Program + manpower support: basic Program + additional CCT visits
- c. Basic Program + SMS support: basic Program + SMS support provided by CCTs

At the district level, SHRP will develop materials to support early grade reading. These materials will be adapted in order to take into account the different needs of learners at different stages of cognitive and academic development, and the linguistic characteristics of the different local languages, rather than be translated directly from one language to another. Furthermore, in order to develop these materials, SHRP will work with MoES and Local Language Boards (LLBs) to standardize orthographies of the target languages. All materials will follow the same general pedagogical framework to facilitate guidelines for textbook development and teacher training.

Finally at the national level, SHRP will work with MoES and the Sector Policy Management Working Group to develop a Uganda-specific reading strategy, which will include policies in the areas of Local Language Board development, textbook development, printing as well as Special Needs. SHRP will also assist the MoES in advocating for reading outside the classroom. Together with MoES, SHRP will aim to raise awareness of local language development, reading instructions and special needs learners by using national communication campaigns through mass media and mobilizing local communities. The national level activities also include a strengthening component of MoES’ ability to monitor reading achievement for research and programmatic purposes.

Result 2: Improved HIV/AIDS Knowledge, Attitudes, and Practices

The Result 2 intervention’s goal is to improve students’ knowledge, attitudes and practices regarding HIV/AIDS. Improving HIV/AIDS education and health supporting attitudes and behaviors will be done by 1) improving MoES planning of the HIV prevention response; 2) improving coordination between MoES and other stakeholders; 3) supporting the school-level impact of HIV/AIDS education; 4) improving the integration of HIV/AIDS Education into MoES Investment Planning and 5) supporting programs and policies with data and research.

At the school-level, this intervention will be implemented in a subset 150 of Result 1 intervention primary schools, 1as well as an additional 50 secondary schools. The target student population of the HIV/AIDS intervention spans grades P4-P7 through S1-S4. Unlike Result 1, teachers of the selected intervention schools will be trained directly rather than via the CCTs. SHRP will support MoES to enhance the Presidential Initiative on AIDS Strategy for Communication to Youth (PIASCY) curriculum, and establish a minimum package of HIV education interventions in addition to providing teacher trainings.

At the district and national levels, SHRP will support the MoES to improve the HIV/AIDS education assessment and reporting system in order to enhance evidence-based monitoring and evaluation. SHRP will also assist MoES to improve coordination between the Ministry and other stakeholders in HIV/AIDS education by developing a cross-sector coordination framework at national and district levels.

¹ At the time of writing this report, the treatment arms are still being developed by RTI. This description is our current understanding of the interventions planned.

B. APPROACH TO PERFORMANCE AND IMPACT EVALUATION

Since the inception of the Performance and Impact Evaluation Project in October 2012, NORC has been working closely with RTI International (RTI) and USAID to design and implement a rigorous impact and performance evaluation of the SHRP Project. The basis of the evaluation designs is the methodologies described in NORC's proposal to USAID/Uganda for the P&IE project. However, as expected, the original designs, which were proposed without detailed knowledge of the realities of SHRP, have undergone modifications based on the realities of project implementation.

In particular, during the past three months, the NORC evaluation team has been working closely with the RTI implementation team refining the impact evaluation design, as appropriate, given implementation realities, and jointly agreeing on key evaluation parameters, which include the following:

- **Evaluation Questions/Hypotheses, and Indicators.** Based on a better understanding of the SHRP interventions and how they will be implemented, RTI's data collection plan and PMP, and a close review of questionnaires, NORC and RTI have reached consensus on a set of impact indicators that will be measured through the impact evaluation. The primary focus of the literacy intervention, Result 1, will be improved literacy as measured by scores in reading comprehension, listening comprehension, vocabulary knowledge, reading fluency and letter sound knowledge. For the school health intervention, Result 2, the focus is largely on improved knowledge and attitudes, with some measures of self-reported behavior change as measured through condom use, age of sexual debut, and number of sexual partners.

The key set of questions for the Performance Evaluation was shared with the SHRP team in November 2012. They are included in this report, but will be refined further as the midterm performance evaluation draws near.

- **Treatment and Control Groups.** The treatment and control groups, as described in NORC's proposal have been largely retained. The final design will comprise treatment and control groups within treatment districts, as well as control schools in matched comparison districts, allowing us to measure the impacts of both school- and district-level interventions.
- **Magnitude of Change and Sample Size.** Sample size for the impact evaluation and the scope of the corresponding data collection depends on the expected magnitude of change of outcome indicators of interest. NORC estimated sample sizes for both Result 1 and Result 2 interventions based on the minimum detectable effect sizes (MDES) posited to us by RTI for the two sets of interventions. Similarly, the time it takes to detect changes in outcomes of interest (literacy skills, attitude change) determine the timing of data collection. NORC has been working with RTI to define a data collection plan that is compatible with the impact evaluation design and within RTI budget, and presented this plan to USAID for review and approval.
- **Data Requirements.** RTI is responsible for collecting the data necessary for the impact evaluation. NORC has been working closely with the SHRP Monitoring & Evaluation and technical teams to reach consensus on data collection needs, including questionnaire content, sample size, and timing and frequency of survey rounds. The final agreements on data collection are presented in this report, and have also been submitted separately to USAID for approval. To ensure that data is of high quality, the NORC team is working with the implementer on instrument review and testing, enumerator training, field procedures and data quality assessment. We will also solicit feedback from USAID, MOES, and RTI on key informants and data sources for the performance evaluation.

C. PERFORMANCE EVALUATION

Performance Evaluations (PE) of the SHRP Project will take place in August 2014 and March 2016 as mid-term and end-of-project evaluations. The purpose of the performance evaluations is to provide rich qualitative data on program design, implementation and effectiveness in order to investigate how these relate to the quantitative data on student learning that will be collected through MOES-led measurement systems and an external impact evaluation. The PE will also serve to provide valuable information on the strengths and weaknesses of program implementation that can be used to improve the design and implementation of SHRP interventions during scale-up, and/or to inform the implementation of similar projects in other locations and contexts.

NORC will design and implement the performance evaluation using best practices in evaluation, including:

- Using subject matter specialists in literacy and health
- Obtaining feedback from the implementing team on the PE methodology and questions
- Using a combination of qualitative and quantitative information
- Fostering transparency, adaptation, and learning by disseminating the PE findings to key stakeholders, most importantly, the implementing team and USAID

Although the bulk of the performance evaluation will occur in 2014 and 2016, during which time, NORC's PE team will travel to Uganda to conduct structured key informant interviews and observational field visits, the collection of data and information for the evaluation is an ongoing process. NORC's local staff has attended all SHRP workshops and meetings for Results 1 and 2 in December 2012 and January 2013, as a means of observing and gathering information on project implementation and processes. This close involvement of evaluation staff in the SHRP implementation process will continue during the actual implementation of literacy and school health interventions at the school-level in 2013. As such, the performance evaluation will be an underlying activity that spans the entire implementation timeframe, accentuated by two structured data gathering and analysis efforts in August 2014 and March 2016.

C.1 Approach

The performance evaluation will assess program effectiveness, and achievement against planned five-year SHRP results.

To assess **program effectiveness**, the performance evaluation will:

- Assess the extent to which the program components are achieving stated goals and objectives as measured against key program documents such as the cooperative agreement, results framework, work plans, and Performance Management Plan (PMP).
- Provide an understanding of progress by program rationale, impact, cost-effectiveness, and sustainability (engagement and ownership)
- Identify if there are management, coordination, and implementation practices that need to be maintained, stepped up, modified, or discontinued
- Consolidate lessons learned and best practices to promote scale up in this important and innovative area
- Assess the validity of the development hypotheses:

Result 1: By focusing interventions on the nexus of language, pedagogy, and instructional materials, USAID can significantly improve students' early grade reading and P3 literacy scores within targeted schools and districts.

Result 2: By strengthening cross-sector coordination between USAID's health and education partners, USAID can significantly improve teachers' and students' HIV/AIDS knowledge and skills within targeted schools and districts.

To assess **achievement of planned five-year SHRP results**, the performance evaluation will examine the degree to which the following process and outcome results have been achieved:

1. National policy framework and Thematic Curriculum enhanced to strengthen the pedagogical framework early grade reading and transition to English
2. At least 3.5 million children demonstrating improved reading skills over the baseline levels for those grade levels
3. At least 10% of P2 students in target schools and districts demonstrating sufficient reading fluency and comprehension to 'read to learn'
4. 65% of students meeting Uganda's national literacy standards by P3 (NAPE)
5. 55% of students meeting Uganda's national literacy standards by P6 (NAPE)
6. Equity improved across genders, geographic regions and languages in early grade reading fluency, and in literacy at the P3 level (NAPE)
7. Language-based instructional materials developed for teachers and students to support the P1-P4 thematic curriculum and promote a reading culture
8. HIV/AIDS education assessment and reporting integrated into MOES systems
9. Cross-sector health and education coordination on HIV/AIDS and health strengthened at the national, district, and school levels
10. Improved HIV/AIDS and health knowledge demonstrated by teachers and students in target districts over the baseline levels for target group

C.2 Data Sources

Data for the performance evaluation will come from the following sources:

- Review of program documents and implementation materials such as teaching materials
- Participation and observation of key implementation activities including, but not limited to, workshops and work meetings, training sessions, school-level interventions
- Gathering of qualitative information through:
 - ✓ Key informant interviews and focus group discussions with stakeholders, implementing partners and beneficiaries
 - ✓ School visits and classroom observation
- Quantitative data:
 - ✓ RTI survey data and P&IE annual impact evaluations
 - ✓ SHRP EGR assessments

- ✓ Uganda National Examination Board data on student performance

The performance evaluation will consist of an analysis of the data from these multiple sources determine performance effectiveness and results achievement, and developing findings and conclusions.

Below, we present our framework for assessing program effectiveness and planned program results.

Table 1.A Performance Evaluation Framework for Assessing Program Effectiveness

	Performance evaluation questions	Data sources
Rationale	<p>What is the defining rationale for the strategies and activities implemented under the Literacy and Health Education Program</p> <p>What priorities guide the program? How have they been identified? By whom?</p> <p>Have the program’s strategic priorities been effectively translated into a clear, coherent, focused plan of support to the MOES, target districts and schools?</p>	<p>Program Cooperative Agreement and Amendments</p> <p>Work plan and Gantt Chart</p> <p>Program Monitoring Plan</p> <p>Program reports: quarterly and annual reports, trip reports, ad hoc reports and presentations, meeting notes</p> <p>Ministry/USAID/Project Strategy Documents</p> <p>Key informant interviews (KIIs) with program implementers, USAID, PEPFAR, MOES, MOH, UAC, NGOs, Language Boards, and other donors</p>
Implementation	<p>Is the program meeting its deliverables and targets for each result indicator?</p>	<p>Program Monitoring Plan</p> <p>Monitoring data and program reports</p>
Impact	<p>Rate how well each program component is contributing to the program and/or assistance objectives.</p> <p>Which interventions have the greatest effect on reading skills acquisition? Which have the least?</p> <p>Are some program components having better success in some schools (context) than others?</p> <p>What are the key factors for the differences in performance in some schools (contexts) receiving the same interventions?</p> <p>At what point (reading stage, grade) are students making the transition from learning to read to reading to learn?</p>	<p>KIIs with selected gov’t officials within MOES, MOH, UAC, and Language Boards at the district and national level. District level key informants include district education officer (DEO) and district health officer (DHO), HIV/AIDS focal persons, and community development officers (CDOs).</p> <p>KIIs with USAID and implementing agency staff, including sr. managers, technical advisors, and M&E staff.</p> <p>Focus group discussions (FGDs) with students, parents, and teachers in intervention districts. KIIs with school administrators, including head teachers, CCTs, and PTCs.</p> <p>KIIs with local NGOs/CBOs and community based trainers (CBTs) involved in HIV/AIDS education.</p> <p>Classroom observation and review of classroom instruction materials.</p> <p>Review of policy documents and curricula.</p> <p>P&IE survey data and annual impact evaluations</p> <p>SHRP EGR assessments</p>

<p style="writing-mode: vertical-rl; transform: rotate(180deg);">Sustainability</p>	<p>What is the sustainability plan? What are the factors contributing to sustainability?</p> <p>What level of engagement and ownership is demonstrated by the MOES and other stakeholders of the program? What are their perceptions of the program?</p> <p>To what extent will the programs components and subcomponents continue without USAID assistance?</p> <p>How can components become more sustainable?</p> <p>What resources (e.g., instructional materials) have been produced? How have they been developed/distributed? With what result? Are they being used? What is the plan for continued availability?</p>	
<p style="writing-mode: vertical-rl; transform: rotate(180deg);">Cost-effectiveness</p>	<p>What are the costs and impact associated with the strategic approaches, activities and “treatments”?</p> <p>What are the implications and recommendations for potential scale-up of program interventions?</p> <p>In what ways can the programs be more cost effective?</p> <p>Are there costs that can be absorbed by the government, community, school budget, or private sector?</p> <p>Are there ways the reading books could be locally produced and distributed (if applicable)?</p>	<p>Budget data</p> <p>KIIs and FGDs indicated above</p> <p>P&IE survey data and annual impact evaluations</p> <p>SHRP EGR assessments</p>
<p style="writing-mode: vertical-rl; transform: rotate(180deg);">Management/ Coordination/Lessons Learned</p>	<p>How can program design, management and execution become more efficient toward achieving program goals?</p> <p>Has program management been efficient and effective? Are there adjustments to strengthen management?</p> <p>What success has there been in building synergies and leveraging comparative advantages of different partners?</p> <p>How has the MOES and/or any other partners contributed to the funding and implementation of this program?</p> <p>What issues, problems or setbacks have been encountered, and how have they been addressed? What issues/problems are still to be addressed?</p> <p>What opportunities are there to further strengthen the program?</p> <p>What are the unintended consequences/spillover effects?</p> <p>How can the program leverage the unintended positive consequences of the program?</p>	<p>KIIs with selected national and district-level MOES, MOH, and UAC officials and Language Boards. District level key informants include district education officer (DEO) and district health officer (DHO), HIV/AIDS focal persons, and community development officers (CDOs).</p> <p>KIIs with program implementers and USAID staff</p> <p>FGDs with students, parents, and teachers in intervention districts.</p> <p>KIIs with school administrators, including head teachers, CCTs, and PTCs</p>

Table 1.B Performance Evaluation Framework for Assessing Planned Program Results

Program Five-Year Results	Data sources
Process Results	
National policy framework and Thematic Curriculum enhanced to strengthen the pedagogical framework early grade reading and transition to English (#1)	Review of national policy framework and Thematic Curriculum KIs with MOES and RTI
Language-based, instructional materials developed for teachers and students to support the P1-P4 thematic curriculum and promote a reading culture (#7)	Review of instructional materials and printing and distribution records
HIV/AIDS education assessment and reporting integrated into MOES systems (#8)	District reports for the national EMIS
Cross-sector health and education coordination on HIV/AIDS and health strengthened at the national, district, and school levels (#9)	Coordination meeting notes; planning/budget document references
Outcome Results	
At least 3.5 million children demonstrating improved reading skills over the baseline levels for those grade levels (#2)	Number of children reached combined with impact estimates
65% or more of students meeting Uganda's national literacy standards by P3 as defined by NAPE (#4) 55% or more students meeting Uganda's national literacy standards by P6 as defined by NAPE (#5)	P3 National Literacy Exam Scores P6 National Literacy Exam Scores
Equity improved across genders, geographic regions, and languages in early grade reading fluency, and in literacy at the P3 level (NAPE) (#6)	P3 National Literacy Exam Scores disaggregated by gender, geographic region and language (NAPE)
At least 10% of P2 students in target schools and districts demonstrating sufficient reading fluency and comprehension to "read to learn" (#3)	Impact evaluation results EGRA scores
Improved HIV/AIDS and health knowledge demonstrated by teachers and students in □target districts over the baseline levels for target group (#10)	Impact evaluation results Comparison of HIV/AIDS and health knowledge baseline, mid-term, and final scores

D. IMPACT EVALUATION

The purpose of the impact evaluation is to test the program’s development hypotheses by demonstrating the existence or absence of a causal relationship between program interventions and changes in students’ reading skills and knowledge, attitudes and practices regarding HIV and AIDS.

NORC has been working with the program implementer, RTI, to adapt the evaluation methodology presented in our proposal to USAID/Uganda such that it reflects the realities of program implementation.

There were several programmatic and budget considerations that NORC learned of following the award of the contract that led to modifications in evaluation design and data collection plan:

- The early grade reading intervention (Result 1) is being implemented through Coordinating Center Tutors (CCTs) who are each responsible for a certain number of schools, which affects the assignment of schools in treatment districts into treatment and control groups. Under this approach, SHRP trains CCTs in mother-tongue based EGR teaching methods, and those CCTs, in turn, serve as trainers for teachers in their cluster of schools. Given this approach, randomization needed to occur at the CCT level, rather than the school-level as originally envisioned in our proposal.
- The early grade reading intervention has three treatment arms: as described above, the basic program (arm 1) is comprised of training and materials for teachers, with a certain number of follow-up visits by CCTs to provide support and additional guidance; the basic intervention, with additional follow-up visits by CCTs (arm 2 or basic program + manpower support); and the basic intervention, with supplemental follow up taking the form of SMS messages from CCTs to teachers (arm 3, or basic program + SMS support). The exact design of the three development arms is still under development. The evaluation is designed to measure the marginal impacts of each treatment arms.
- A parallel early grade reading effort occurring in some of the same districts – the Mango Tree Learning Initiative – was overlapping in 10 SHRP intervention schools. These 10 schools were eliminated from both the SHRP program and the evaluation sample frame; this could introduce bias to our sample.
- The school health intervention (Result 2) is directed at both primary and post-primary schools, with post-primary institutions divided into secondary schools and Business, Technical and Vocational Education and Training (BTJET) institutions. The impact evaluation takes all three levels of educational entities into consideration.
- Due to budget constraints, we streamlined the originally proposed rounds and levels of data collection. These decisions are described in detail in Section D4

Each of these factors is taken into consideration in the impact evaluation design presented below.

D.1 Impact Evaluation Design

NORC is using a combination of an experimental (randomized controlled trial, or RCT) and quasi-experimental (matched comparisons) design to detect the impacts of the School Health and Reading Program. This mixed-method design allows us to estimate the combined effects of the district- and school-level interventions that comprise the School Health and Reading Program; it also allows us to isolate the effects of the school-level intervention for the literacy component of the program.

An impact evaluation (IE) is done to assess the causal effect of a specific intervention on a set of outcomes. It allows us to attribute changes in an outcome to a specific intervention or set of interventions by answering the counterfactual question “What would have happened to program participants in the absence of the intervention?” Ideally, this is done by observing the same program participants both with and without the intervention at the same point in time. Of course, this is not possible; at any given time, a participant either receives the intervention or not. Therefore, we can never directly observe the counterfactual and instead need to create a comparison group to serve as the counterfactual. Identifying a credible comparison group is a critical aspect of an impact evaluation.

The ideal comparison group stems from the use of experimental methods in which eligible participants are randomly assigned to receive the intervention or not. Randomization ensures that, on average, characteristics of the treatment and control groups are statistically identical, with the only difference being their participation in the intervention. In this case, any measured difference in outcomes between

the groups over time can be attributed to the program. When random assignment is not possible, quasi-experimental methods, such as statistical matching, are used to establish a comparison group.

Our impact evaluation design uses both the random assignment of schools to treatment and control groups within SHRP intervention districts (experimental design) and the selection of matched comparison districts (quasi-experimental) in which SHRP is not operating. As discussed below, the experimental design allows us to isolate the effect of school-level interventions from district-level interventions, while the inclusion of non-intervention districts design allow us to measure the impact of the district level interventions and the combined district-school level intervention package.

D.1.1 District-level Comparison Group

In the first year of the program RTI will implement literacy interventions in 11 districts located in 4 different language areas (Table 2). These 11 districts were chosen by RTI and MOES, and were not part of the evaluation design. Therefore randomization at the district level was not possible. However, we were able to select comparison districts that are similar in key characteristics to the treatment districts. Although we had intended to pair a control district to each treatment district, budget and logistical restrictions expressed by RTI resulted in the selection of only one district per language adding to a total of four comparison districts to compare against 11 treatment districts

Table 2 Treatment and Control Districts

Region	Language Area	Intervention District	Control District
Central	Luganda	Gomba Wakiso	Buikwe
East	Ateso	Katakwi Kumi Serere	Ngora
North	Leblango	Apac Lira Kole	Otuke
South West	Runyankore/ Rukiga	Bushenyi Kiruhura Kabale	Ibanda

The four control districts were selected by matching non-intervention and intervention districts in a specific language area according to district characteristics such as NAPE 2011 results on P3 proficiency in oral reading, P3 proficiency in literacy in English, and P6 proficiency in literacy.² Because we were matching only one comparison district to more than one intervention district, we computed a weighted average of treatment districts’ proficiency scores, where the weights are proportional to the number of schools participating in the program during the first year. Through this matching process, we selected four control districts - Buikwe, Ngora, Otuke, and Ibanda (Table 1).

Figure 1 below shows that 11 districts (in green) will receive treatment while 4 districts (in red) will be used as comparison districts.

² Unfortunately, no information about HIV and AIDS knowledge, attitude and practices was available at the time of matching districts.

D.1.2 Assignment to Treatment and Control Groups and Measurement of Impact

Reading Program

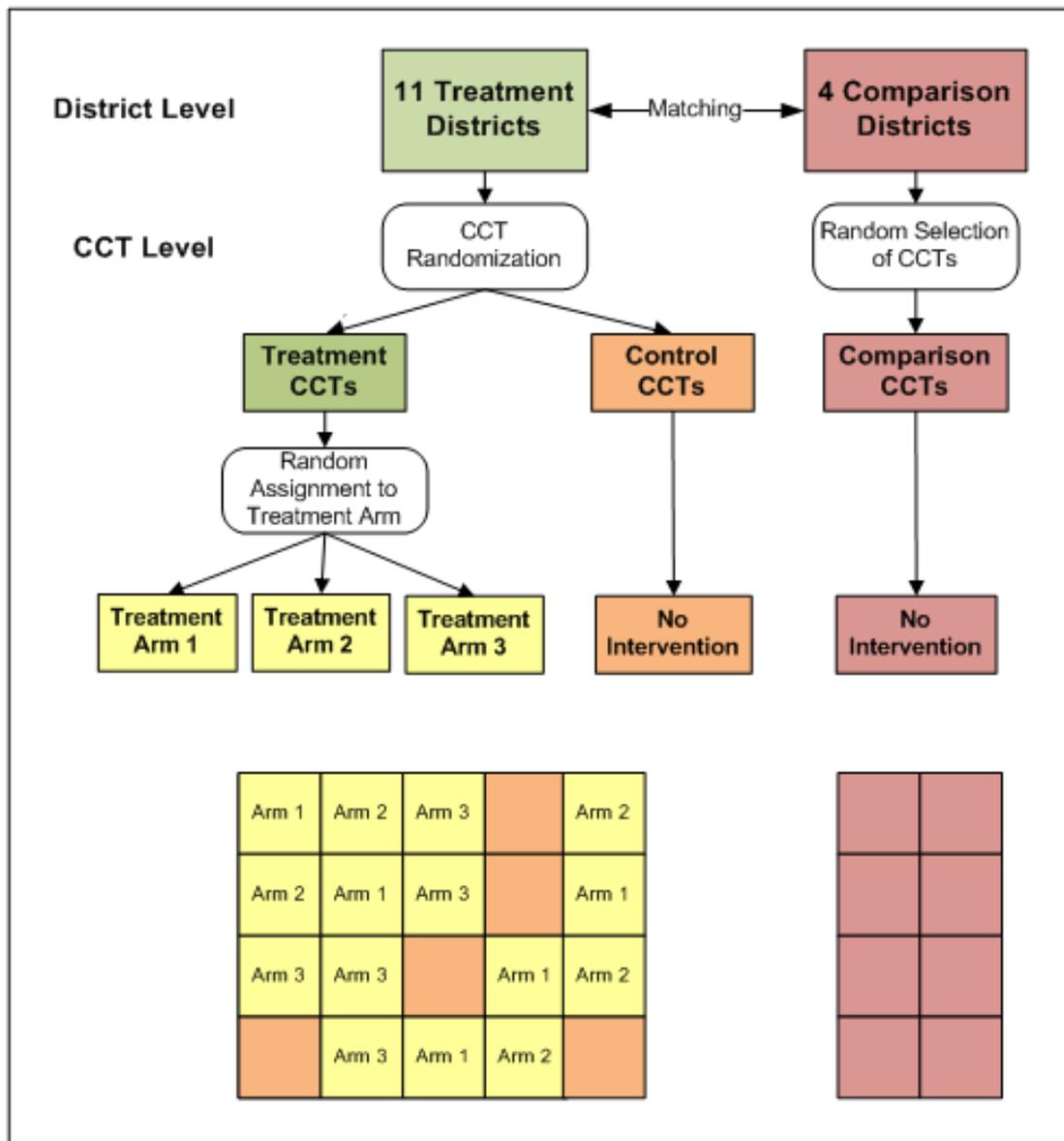
As explained above, the CCTs are important actors in the implementation of the Early Grade Reading Program (Result 1) interventions. They are responsible for training teachers and providing follow-up support and assistance in implementing the Result 1 interventions. Towards this end, SHRP conducted training workshops (training of trainer workshops) for CCTs in different regions. Because of the critical role of CCTs in the implementation of the EGR intervention, and because each CCT is responsible for a cluster of schools, randomization of schools into treatment and control groups has to occur at the CCT cluster level, rather than the school level. Since each CCT is responsible for several schools, randomizing at the school level would imply that a CCT would have to treat schools under his or her jurisdiction differently if some were designated as treatment schools and others as controls. After consulting with RTI's local staff, we reached agreement that this was an unrealistic expectation and that randomization at the school level had a high risk of 'contamination' or 'bleeding' between different treatment arms and between treatment and controls arms. Instead we opted to randomize at the CCT level, assigning the entire cluster of schools under a CCT to either one of the three treatment arms or to the control group. Therefore, all language areas have CCTs assigned to the four possible groups.

In comparison districts we randomly selected CCTs whose school clusters will serve as out-of-intervention-district controls.

Figure 1 shows how, within treatment districts, CCTs are assigned to each arm of the intervention (yellow cells) or to the control group (orange cell), and how, in the comparison districts, some CCTs are selected as controls. This process creates five groups of schools that will be used in the impact evaluation

1. Basic program (treatment arm 1)
2. Basic program + manpower support (treatment arm 2)
3. Basic program + SMS support (treatment arm 3)
4. Controls within treatment districts
5. Controls in comparison districts

Figure 1: Assignment to Treatment and Control at the District and CCT levels for Early Grade Reading Impact Evaluation



Intuitively, the difference in outcome indicators between yellow schools and orange schools will show the effect of the school-level intervention, given that both types receive the district intervention but only yellow schools receive the school-level program. The difference in outcomes between orange schools and red schools will identify the effect of the district-level intervention. While none of those schools benefits from the school level programs, the orange schools are exposed to the district level treatment. The impact of the complete intervention package can be measured by estimating the difference in outcomes between yellow and red schools.

A simple representation of the measure of impact under the experimental and quasi-experimental design is the interaction effect of treatment and time, or the double-difference estimate:

$$Estimate\ of\ impact = (Y_{T,t2} - Y_{T,t1}) - (Y_{C,t2} - Y_{C,t1})$$

where,

- Y = impact indicator
- T = treatment group
- C = comparison group
- t1 = baseline or beginning of study
- t2 = end of study

School Health Program

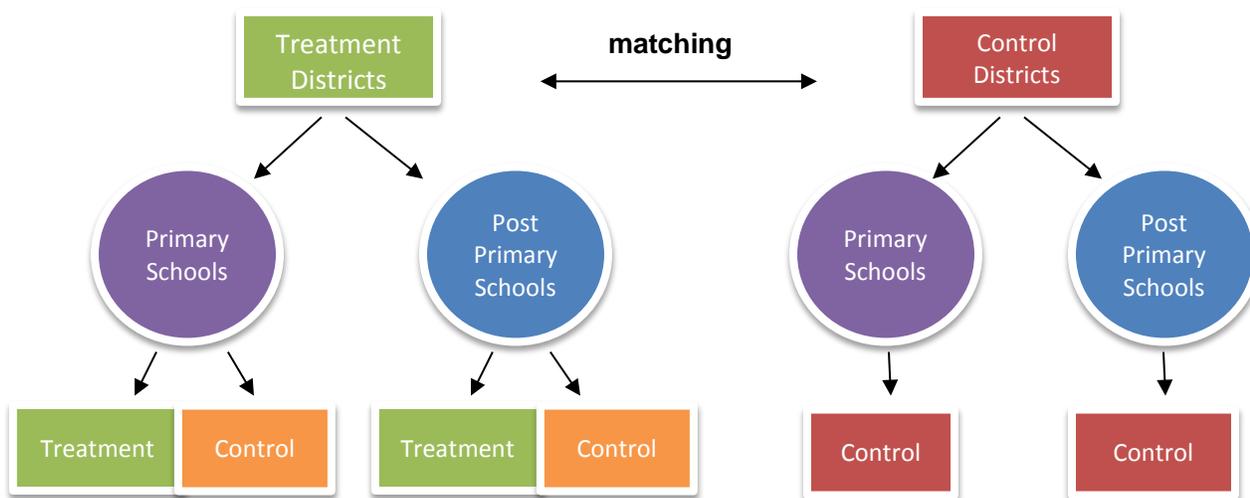
The School Health Program requires a simpler randomization process. First, CCTs are not a factor in this set of interventions. A core group of trainers will directly train teachers and school counselors, who in turn will provide information and counseling students. Second, this intervention has only one treatment arm. Therefore, for the impact evaluation, schools that already receive treatment under the Early Grade Reading Program will be randomly assigned to either the treatment group which will receive the school-level health interventions or to the control group which will not receive any school-level health interventions. . This results in three groups of schools for the impact evaluation:

1. Treatment group
2. Controls within treatment districts
3. Controls in comparison districts

In addition to work in primary schools, the HIV and AIDS program will be rolled out in post primary schools.

NORC randomly selected the schools from the population of primary and post primary schools in the target areas.

Figure 2



As in the case of the Reading Program, we will compare the evolution of the outcome indicator over time in the treatment schools versus the control schools. The difference in outcome indicators between green schools and orange schools will show the effect of the school-level intervention, given that both types receive the district intervention but only green schools receive the school-level program. The difference in outcomes between orange schools and red schools will identify the effect of the district-level intervention. While none of those schools benefits from the school level programs, the orange

schools are exposed to the district level treatment. The impact of the complete intervention package can be measured by estimating the difference in outcomes between yellow and red schools.

D.1.3 Estimating Impact

We will estimate the impact of the activities in several different ways. We will start with the simplest analysis, which is the estimation of the *impact of the interventions within one year of school*. Suppose for example that we want to evaluate the effect of the Reading Program on fluency scores, Y , for students in grade P1 in the Luganda speaking areas. We will use the scores collected at the beginning of the school year 2013 (baseline) and from the end of the same school year (endline), and regress the change in the test score Y of student i on the treatment status of the school s and district d ,

$$Y_{isdE} - Y_{isdB} = \alpha + T_{sd} \beta + D_d \gamma + X_i \delta + \varepsilon_{isd}$$

where X_i are individual characteristics of the student i , such as age and sex, T_{sd} is a dummy equal to 1 if school s in district d received the intervention and 0 otherwise, D_d is a dummy variable equal to 1 if the district received the intervention; B indicates "beginning of the year" and E indicates "end of the year." The parameter γ is the effect of the district-level intervention and $\beta + \gamma$ is the effect of the full treatment, i.e. school and district interventions. We can also add controls for the beginning of the year test score of the student. Given that test scores tend to have a strong persistent component, their inclusion increases the precision of the estimated effects.

A similar analysis can be performed for the School Health intervention. In this case, the program is the same for all students P4-P7 and for all post-secondary students and therefore the analysis may be done by grade or controlling for grade/age. Decisions in this respect are pending. Initially the idea proposed by the RTI School health team was to collect enough data to perform the analysis by grade; however budget restrictions may require reducing the sample size.

The advantage of estimating the effect of the intervention within one year/grade is that we can do this immediately after the first year of deployment of the program and have some initial indication of its effects. This will be particularly useful to inform the differences between the different intervention arms in the Reading Program. However, the effects of the program could be greater if the students are exposed to the intervention over time, across different grades. Therefore, we also propose to estimate the *cumulative impact of the intervention* for a given cohort of students.

Figure 3 shows the cumulative effect of the intervention on students depending on the grade they are attending at the start of the program and the specific year of program implementation. For instance after 4 years of the program (end of 2016), Cohort 1 (in figure 2) will have received four years of intervention, Cohort 2 three years and Cohort 3 two years. After four years of intervention, we will therefore be able to calculate the difference in P3 tests scores between treatment and control schools for Cohort 1 students exposed to the program since P1 (for three years), and calculate this difference in P2 test scores for Cohort 2 students exposed to the program for 2 years and P1 test scores for Cohort 3 students exposed to the program for one year. In scientific notation, this evaluation model may be represented as follows:

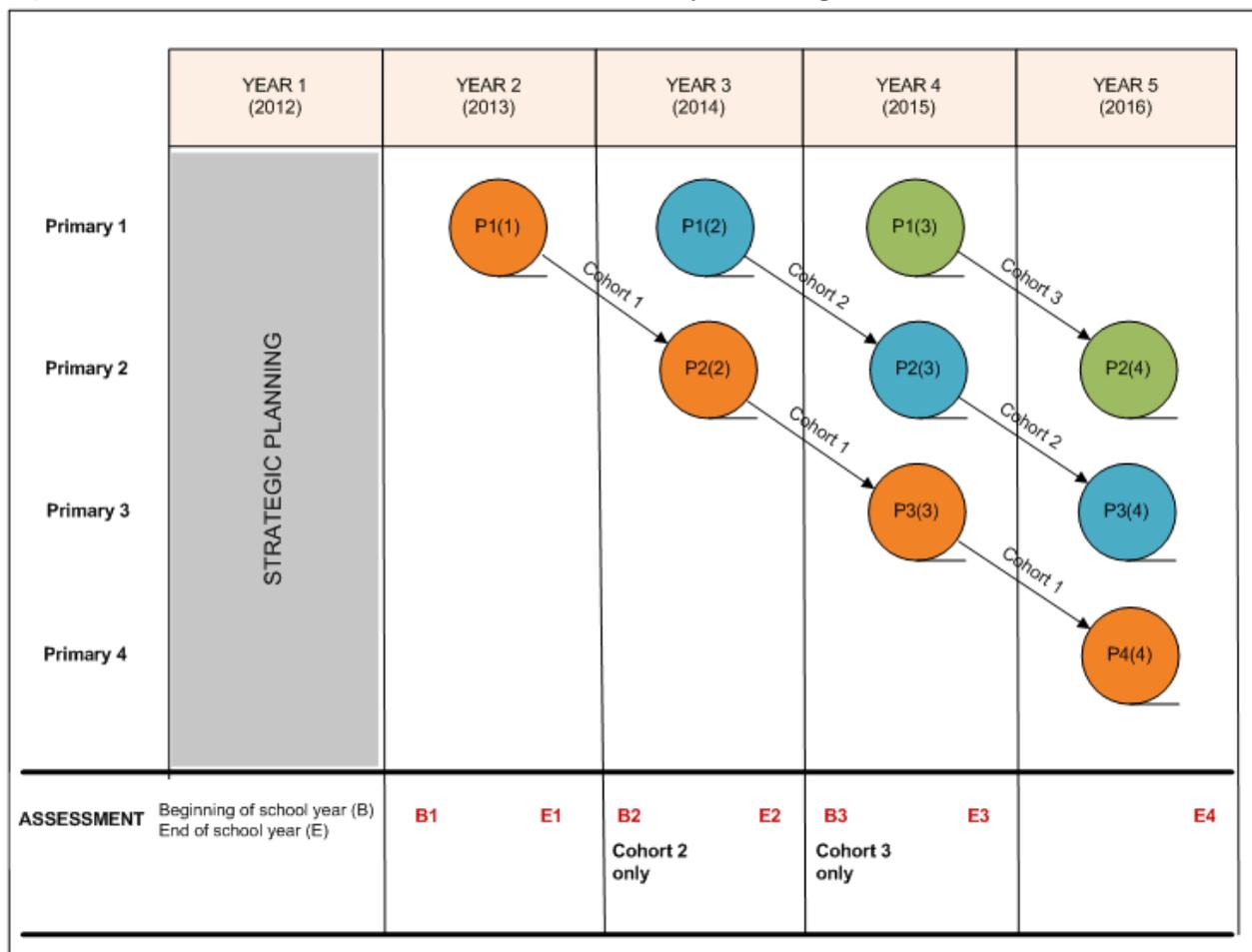
$$Y_{i3sE} = \alpha + \sum_{g=1}^3 T_{igs} \beta_g + X_i \delta + \varepsilon_{igs}$$

As before T indicates treatment; the sum allows to account for exposure to treatment in P1, P2, P3 or never.

A similar analysis may be done for the HIV/AIDS interventions. Following cohorts across grades, it may be possible to estimate the cumulative effects of exposure to health education over time. Alternatively, we can estimate the impact effects within a year as explained above.

Finally, for the literacy interventions, we can also explore whether there are *teacher effects*. Teachers may become more effective over time as they receive more training or gain experience teaching the new curricula and using the new materials³. This means that teachers could be more effective in Year 4 of the program than at its very beginning. In other words, in Figure 2, Grade P1 students in Year 4 of the program (P1(3) in green color) could receive better instruction than Grade P1 students in Year 2 (P1(1) in orange color). The impact of the teacher effect may also be estimated if it is of interest to USAID/Uganda and the MOES.

Figure 3: Illustration of Grades and Cohorts, Early Reading



³ It has been suggested by Benjamin Piper of RTI that teachers in Uganda do not teach in the same grade over time but rather moved with their cohort of students. If this is the case, it may complicate the estimate of teacher effects however it has not been confirmed yet.

D.2 Impact Indicators and Data Sources

D.2.1 Result 1 – Reading Program

Literacy is comprised of multiple skills, both receptive and productive. Successful readers must be able to identify letters and their corresponding sounds, segment and blend those sounds to form words and sentences, master appropriate vocabulary, and make meaning from text, among other skills. They must also be able to demonstrate their understanding and engagement with text through writing. To assess the effectiveness of the SHRP in reaching its goal to improve early grade reading and transition to English, specific key literacy skills will be assessed.

The consensus among the reading research community in the United States is that effective reading instruction attends to *at least* five main reading skill areas including alphabetics (letter knowledge and phonemic awareness), fluency, vocabulary, and comprehension and phonics (National Reading Panel, 2000; Snow, Burns & Griffin, 1998).⁴ Based on this research, the Early Grade Reading Assessment (EGRA) a brief oral reading assessment that tests these skills, will be used to measure program impacts on literacy (RTI International, 2007).

EGRA is comprised of multiple sub-tests that focus on the five main reading skill areas outlined above. Within each of these five areas, there are multiple sub-tests that can be selected for inclusion, based on local needs and the goals of the assessment system. Table 3, below, outlines some of the most common early literacy skill sub-tests that are included in the various iterations of EGRA.

Table 3: Early Literacy Skill, Sub-test

Early Literacy Skill	Sub-test	Measurement
Alphabetic Knowledge	Letter Sound Knowledge	Number of letter sounds correctly identified out of 100 in 60 seconds
Phonemic Awareness	Segmenting	Number of phonemes correctly identified out of the total number found in 10 words (exact words and number of phonemes to be determined)
Phonics/Alphabetic Principal	Nonword decoding	Number of nonwords correctly decoded out of 50 in 60 seconds
Fluency	Oral passage reading	Number of words in a reading passage of approximately 68 words read fluently (with accuracy) in 60 seconds
Reading Comprehension	Oral recall	Number of questions (out of four) about a reading passage (read by student) answered correctly
Listening Comprehension	Oral recall	Number of questions (out of four) about an passage read aloud (by facilitator) answered correctly
Vocabulary	Oral identification of common objects	Number of common objects correctly identified

No clear benchmarks for the EGRA tool have been established. That is, the EGRA tool provides a snapshot of early literacy skills but does not provide guidelines for interpreting which children can be

⁴ These five skills are not meant to be all inclusive; however, considerable empirical research has been conducted in these skill areas that has indicated they are important predictors of reading.

considered “readers” or what level of performance should be expected on each sub-test. At the same time, EGRA has been used to assess early literacy skills in more than 50 countries around the world; thus, performance of students participating in SHRP can be compared with the range of performance of other children on EGRA in other low-income countries.

A notable component of SHRP is its transitional bilingual design. That is, literacy instruction will begin in one of four mother tongue languages, with English introduced as a subject area nearly simultaneously (within 4-8 weeks after mother tongue instruction has begun). The language of instruction will then increasingly transition from mother tongue to English over the course of four years. Because of this transitional bilingual design, the impact evaluation necessarily requires a heteroglossic⁵ approach to assessment. Early literacy skills will be assessed in mother tongue and English and the relationships between literacy skills in L1 and L2 will be examined.

This transitional bilingual design has implications for the sub-tests that will be included in the EGRA tool in each language. For example, because most grade 1 students cannot be expected to have prior knowledge in English language or literacy, the sub-tests that have been selected to assess literacy in English are aimed at capturing lower skill levels; in contrast, students are expected to already possess basic linguistic knowledge in their mother tongues and the EGRA sub-tests that have been selected aim to capture a distribution of literacy skills that include higher level abilities. For these reasons, the initial baseline tools across languages will vary slightly. Table 4 below illustrates the sub-tests that will be included in English and mother tongue EGRA.

Table 4: Sub-tests for English and Mother Tongue EGRA at Baseline

Sub-Test	English	Mother Tongue
Letter –sound knowledge	X	X
Phonemic segmenting	X	X
Non-word decoding	X	X
Receptive vocabulary	X	--
Oral reading fluency	--	X
Reading comprehension	--	X
Listening comprehension	--	X

The recommendation of the NORC evaluation team to RTI was to consider adding some additional sub-tests of higher difficulty to future rounds of the English assessment, to gauge expected student improvement over time. Likewise, the NORC team has encouraged RTI to consider developing more sophisticated vocabulary and writing sub-tests, to provide a more holistic understanding of the impact of SHRP on student literacy acquisition in all languages.

Secondary impacts on students and teachers may also be expected. Impacts on students might range from higher rates of school attendance in the short term, to lower drop-out rates in the long term. For example, we hypothesize that with improved instruction in their mother-tongue, students will be more engaged in the classroom, demonstrate greater learning and will be less likely to miss school. Higher attendance rate and thus higher exposure to literacy instruction will also contribute to improved literacy. All of these factors (higher attendance, improved literacy) may also help keep the students in school longer over time, thus decreasing the drop-out rate⁷.

Secondary impacts on teachers might range from higher rates of school attendance to more effective use

⁵ A “heteroglossic” approach conceptualizes literacy learning in both mother tongue and English as interconnected, co-existing, and mutually reinforcing.

of classroom time during reading instruction. High rates of teacher absenteeism, ranging from 27 to 43%, have been identified in studies of Uganda (Rogers & Vegas, 2009; Yiga & Wandega, 2010). While numerous factors beyond the control of SHRP are likely contribute to absenteeism, we hypothesize that increased professionalism and teacher efficacy resulting from high quality professional development will likely increase teacher attendance. Likewise, we expect the professional development intervention to first improve teacher knowledge and practice, as a means to improve student literacy outcomes. Thus, we expect that teachers will demonstrate higher quality literacy practices, as reported by students. An expanded list of potential secondary outcomes is as follows:

Table 5: Expected secondary outcomes and their measurements

Level	Indicator	Measurement	Expected Outcomes
Student	Attendance	Is there any day you did not come to school last week?	Decrease in reported absences
	Motivation to Read	Do you like to read?	Increase in reported motivation to read.
	Self-efficacy	Do you consider yourself a good reader?	Increased self-efficacy
Teacher	Attendance	Is there any day your teacher did not come to school last week?	Decrease in reported absences
	Literacy Instruction	Does your teacher use special materials when he or she is teaching reading?	Increase in access to materials during reading lesson
		During the literacy hour do you ever: <ul style="list-style-type: none"> Learn about the sounds that letters make Copy words into your exercise book Answer questions about a story the teacher has read to you Write your own stories 	Increase in effective literacy instruction activities
Materials	Use	Do you bring home reading books from your classroom or from the school library to read at home?	Increase in taking books home
	Language	What language are these books or materials in?	Increase in access to mother tongue materials
Outside of School Literacy Practices	Reading	Does anyone at home read to you? ⁶	Increase in reported read aloud at home
		Do you see anyone in your home read newspapers, religious texts or books?	Increase in reported observations of home reading

⁶ This indicator may be measured by frequency, type of reading, and motivation to read as well. For example, children may be asked with what frequency they engage in different types of reading activities outside of school (e.g. read aloud, listen to someone read aloud, talk about books, read for fun, etc.). They may also be asked to what extent they agree with questions such as: “I enjoy reading” or “Reading is boring”.

These potential secondary outcomes will be measured through a learner context questionnaire that will be administered to the student at the same time as EGRA, thus allowing us to conduct analysis to determine if the EGR intervention has had an effect of these outcomes. The learner context questionnaire also includes questions about students’ demographic and socio-economic characteristics as well as information on initial home literacy environment which can be used as explanatory variables in our analyses.

D.2.2 Result 2 – School Health Program

USAID/Uganda’s health education intervention aims to improve HIV/AIDS knowledge and skills among teachers and students by strengthening cross-sector coordination and systems, defining a minimum standard package of school-level health education interventions, and promoting the availability and use of instructional materials and child-friendly educational activities in schools.

Mainstreaming HIV/AIDS education into the school curricula is directed primarily and most immediately at improving HIV/AIDS-related knowledge and life skills among teachers and students. Similarly to the tool development process for the EGR intervention, NORC has been working with WorldEd (which is implementing the health intervention as part of a consortium with RTI for SHRP) to develop a Knowledge, Attitudes and Practices (KAP) survey to measure key HIV/AIDS indicators. The tool will mostly measure knowledge and attitudes about HIV/AIDS, and perhaps some questions about practices for the older learners. It includes questions related to HIV transmission, HIV prevention and treatment, and related to attitudes towards sexual intercourse.

However, due to delays in implementation and the sensitive nature of the questions related to HIV/AIDS, the Result 2 data collection instrument is still under development. WorldEd has drafted a first version of the tool which has been shared with RTI, NORC and MoES but such an instrument has never been used in Uganda with students as young as those targeted by the intervention (students could be potentially as young as 8 or 9 years old). The tool is awaiting IRB approval and will need to undergo rounds of pre-testing before question items and indicators can be finalized. The final set of indicators will be finalized in close collaboration with RTI, USAID and MoES.

D.3 Sampling

NORC’s statistician and evaluation experts conducted a statistical power analysis to estimate the sample size required for the evaluation of the Reading and School Health Program. The standard approach to determining sample size for analytical surveys is to estimate the sample size required to achieve a specified level of power (probability), such as 90 percent, for detecting a change of a specific magnitude. This sample size depends on a number of factors including the evaluation design, the impact estimate, the design of the sample survey used to collect data, the statistical test, and the population under investigation.

As described above and shown in Figure 1, an initial group of 11 districts located in 4 different language areas was selected by RTI and MoES to participate in the Reading Program. NORC selected a sample of 4 comparison districts. Each comparison district was individually matched on the basis of P3 and P6 NAPE literacy scores to each of the 4 language areas. Within each area CCTs were randomly assigned to 4 arms of the intervention (3 treatment arms and 1 control arm). NORC calculated the number of schools needed in each language/arm cell (20 cells) and within each cell, RTI selected the requisite sample of N treatment schools and N control schools using random assignment, and N_s comparison schools in each comparison district. This “balanced” design is an efficient one, with high return of precision and power for survey resources expended.

A subset of the schools in the Reading Program sample will constitute the sample to be used to evaluate the School Health Program. In addition a sample of post primary schools was created by NORC by randomly selecting from the population of schools in the areas where the Program will take place.

Details about sample size calculations, assumptions and decisions are presented in Annex 1. Below, we present the final sample sizes estimated for the evaluation design. These samples constitute the scope of data collection for the impact evaluation.

For the impact evaluation of the Reading Program, we estimated the sample required to detect a double-difference measure of impact of magnitude $D = 0.20$ with a power of 90%. Based on these calculations for each of 20 cells, it is necessary to have 14 schools, each with 30 P1 students, for a total of 420 students, who will be followed over subsequent years. With 20 cells (3 arms and 2 controls, and 4 language subgroups per group), the total sample size required amounts to 8,400 student in 280 schools; i.e. 8,400 P1 students at baseline in 2013 to be followed in November 2013, November 2014, and November 2015). Of the 280 schools, 168 (5,040 students) would constitute the treatment group, and 112 schools (3,360) would be controls.

Based on these estimations, RTI randomly selected 168 treatment schools for the evaluation sample, from 410 randomly selected intervention schools. Control schools within the treatment districts were selected from the schools in those districts that were not selected for the intervention.

The sample for the School Health Program is comprised of three cells: treatment, controls within the treatment district, and controls in comparison districts. A school sample of size about 234 schools is adequate to detect effects (double-difference impacts) of magnitude $D = 0.20$ with high power (90%). Out of 150 intervention schools, 78 schools are selected for assessment. A similar number of schools is selected for the in-district control and out-district control groups. Additionally, all 50 intervention post-primary schools are included in the evaluation, as well as a similar number of post-primary schools for the in-district and out-district control groups. In order to reduce data collection costs, a total of 30 students per school will be selected for data collection, rather than 30 students per grade. Therefore, for the Health program, a total of 7020 primary school students and 4500 post primary school students will be included for data collection. Table 8 below summarizes the sample sizes for both the Reading program and Health program.

Table 6: Sample sizes for Year 1 of data collection, Result 1 and Result 2

Activity	Treatment Districts				Comparison District		TOTAL	
	Treatment		Control		Control			
	Schools	Pupils	Schools	Pupils	Schools	Pupils	Schools	Pupils
Reading Program	168 primary	5040	56 primary	1680	56 primary	1680	280 primary	8400
	(14 per arm/language cell)		(14 per arm/language cell)		(14 per arm/language cell)			
Health Program	78 primary	2340	78 primary	2340	78 primary	2340	234 primary	7020
	50 post primary	1500	50 post primary	1500	50 post primary	1500	150 post primary	4500

D.4 Data Collection Plan

RTI is responsible for all data collection related to the impact evaluation. As such, NORC has worked closely with the SHRP program statisticians and M&E team to ensure that all instruments follow closely from the evaluation hypotheses and indicators necessary for the impact evaluation. Towards this end, NORC’s subject matter experts and evaluation experts have reviewed all instruments and provided extensive feedback on them to ensure that in addition to outcome indicators related to reading skills and HIV prevention knowledge, the data collection effort includes information on covariates (student’s socioeconomic characteristics, parent education, home literacy environment, etc.) that need to be controlled for in the evaluation model.

Furthermore, NORC’s evaluation expert worked closely with the Results Teams and M&E team to ensure that the data collection covers an adequate sample for the evaluation. Below, we present and describe the final data collection plan that RTI and NORC developed for implementation. This plan meets many, but not all, of the original objectives of the impact evaluation, and fits within the budget and logistical constraints of SHRP.

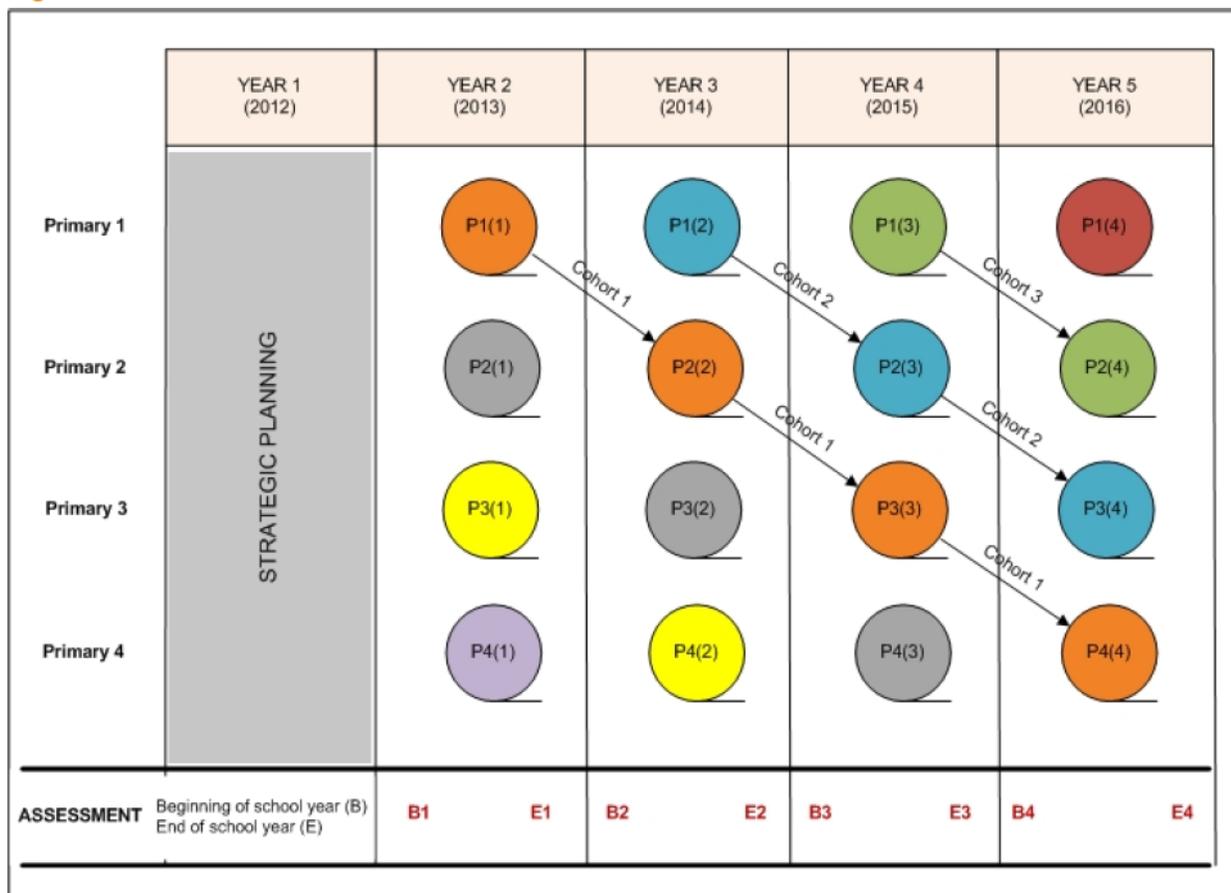
D.4.1 Result 1 – Reading Program

NORC worked closely with RTI to adapt the impact evaluation design in our proposal to the program implementation design and data collection constraints expressed by RTI.

The evaluation plan we delineated in our proposal included the collection of data for four cohorts of students: Cohort 1 (Fig.3, orange circles), Cohort 2 (blue circles), Cohort 3 (green circles) and Cohort 4 (red circle). Our proposed design assumed that these data will be collected from randomly selected treatment and control schools every year for four years and in every district where the intervention will take place. We also planned to collect data from schools in comparison districts. This strategy results in 3 types of schools: those located in intervention districts that receive the program, those located in intervention districts but do not receive the program, and those located in comparison districts. Comparison of these 3 groups allows us to disentangle the impacts of interventions at the school and at the district level, as required per the USAID/Uganda RFP.

Figure 4: Original Plan for Data Collection Proposed by NORC

Figure 2: Illustration of Grades and Cohorts



Based on discussions with RTI, and better understanding of the Reading Program implementation plan and associated budget and time constraints, we are proposing some modifications to data collection assumptions in our original proposal.

Data collection for Cohort 1 (orange circles) will take place as shown in the figure above. RTI has agreed to collect data in treated schools, non-treated schools in intervention districts, and schools in non-intervention districts in 2013, 2014, 2015, and 2016. Per NORC’s sample calculations, which were shared and discussed extensively with RTI over the past 4 months, data collection in the four years will occur in a sample of 30 P1 pupils (over time P1 grade will become P2, P3 and P4) in 14 schools per "cell". As explained in Section 3, the Reading Program will have 20 cells comprising of different combinations of treatment arms or control groups and languages, for Cohort 1.

After 2013 or 2014, RTI may reduce the number of treatment arms, to just one or two. In this case, the total number of schools and pupil required for the sample will be smaller because entire cells will be eliminated. The reduction of treatment arms will be decided by RTI based on the results obtained in the first 2 years.

NORC strongly advised RTI against cutbacks in the data collection efforts for Cohort 1. Following the P1 class in Cohort 1 through P4 is the only way to have a comprehensive evaluation of the impact of the literacy program over the four years for at least 2 languages (maybe 4 languages depending on when materials are ready). This is the only group that receives full treatment from May 2013 until the end of the project. RTI has agreed to our recommendations and will collect data accordingly.

For Cohorts 2 (blue circles) and 3 (green circles), however, RTI would like to reduce the amount of data collected for budgetary and logistical reasons. In response to these constraints, we suggested eliminating data collection in comparison districts. As mentioned above, collecting data in comparison districts allows us to isolate and measure the impact of district level interventions. However, given that it was necessary to make some cuts in data collection to stay within RTI’s budget and other logistical constraints, NORC decided that the impact evaluation would suffer the least with this approach. We will still be able to calculate the impact of school level interventions for Cohorts 2 and 3, but we will not be able to compute district level effects for these two Cohorts.

Finally, RTI will not collect data from Cohort 4 (red circles in Figure 2). NORC’s original idea was to compare the results of P1 students in Year 5 of the program (P1(4) in red color) with P1 students in Year 2 (P1(1) in orange color; i.e. P1 students in the first year of the intervention) in order to explore whether there are teacher effects. Teachers may become more effective over time as they receive more training or gain experience teaching the new curricula and using the new materials. This means that teachers could be more effective in the final year of the program than at its very beginning. On the other hand, some interventions tend to be effective while they are new. After the novelty effect wears off, the positive results may vanish. It is an empirical question as to which effect dominates, and our intent was to explore whether the results for a particular grade change (improve or deteriorate) over time. Given that we will not have test scores for the 4th year of the program, we plan to do the same analysis over a three years of the intervention, using P1 students in Year 4 (P1(3) in green color in Fig. 2). The exercise will be identical but the comparison groups will be closer in time. This means we will have fewer years for the teachers to master the new instruction approach and materials and/ or for the novelty effects to fade away.⁷

Table 4 below shows RTI’s current data collection plan. In bold is the data that will be used in the impact evaluation by NORC. Data collection noted in red font are not required for the impact evaluation, but will be used by RTI for other reporting. For example, RTI is collecting data on P2 to report to MOES and USAID every year.

NORC will report the impact of the project on P2 pupils for every cohort but only after those pupils have been fully treated (i.e. after receiving treatment in P1 and P2).

Table 7: Early Grade Reading Assessment Data Collection Plan: 2013-2016

	2013		2014		2015		2016	
	FEB	NOV	FEB	NOV	FEB	NOV	FEB	NOV
Cohort 1 A (4 LANGUAGES)								
Treatment	P1:30 P3:10	P1:30 P2:10		P2:30		P2 P3:30		P2 P4:30
Control w/in	P1:30	P1:30						

⁷ Please note we are assuming that teachers mostly stay teaching a particular grade (i.e. the P1 teacher instructs P1 every year). Benjamin Piper from RTI stated that in Uganda, teachers do not stay teaching the same grade but move with their cohort of pupils as they progress during primary school. If this is the case, it is likely that the proposed analysis will present some difficulties. We are currently trying to learn how this aspect of the system works.

district				P2:30		P3:30		P4:30
Control out district	P1:30 P3:10	P1:30 P2:10		P2:30		P2 P3:30		P2 P4:30
# of schools	280	280		280		TBD		TBD
Cohort 1B								
Treatment	P1:10			P1:10		P2		P3
# of schools	20			20		20		20
Cohort 2 (8 LANGUAGES)								
Treatment			P1: 30 P3:10	P1: 30 P2:10		P2:30		P2 P3:30
Control w/in district			P1:30 P3:10	P1:30 P2:10		P2:30		P2 P3:30
Control out district								
# of schools			TBD	TBD		TBD		TBD
Cohort 3 (12 LANGUAGES)								
Treatment					P1:30 P3	P1:30 P2		P2:30
Control w/in district					P3 P1:30	P2 P1:30		P2:30
Control out district								
# of schools								
Total # of schools	300	280	TBD	TBD	TBD	TBD		TBD

NOTES: In bold indicates required for impact evaluation. 30=number of pupils per grade/school.

14 schools in each cell (language/arm combination)

In Red denotes that is needed for RTI PMP indicators/USAID reporting and it will not be used for IE. Does not meet sample size requirements as data needed for impact evaluation.

D.4.2 Result 2 – School Health Program

As we write this report, RTI is working on the data collection plans for the School Health Program. The initial intention was to collect data on 30 students per grade in each school in the sample at the beginning and end of each year. Currently RTI is revising this plan, in an attempt to reduce data collection costs.

In order to reduce costs without compromising the rigor of the evaluation NORC is recommending the data collection strategy depicted in Table 5. The strategy is similar to the one used for the EGRA data collection.

Table 8: HIV and AIDS Assessment Data Collection Plan: 2013-2016

	2013		2014		2015		2016	
	FEB	NOV	FEB	NOV	FEB	NOV	FEB	NOV
Cohort 1 (4 LANGUAGES)								
Treatment	P4-P7 S1-S5	P4-P7 S1-S5		P5-P7 S1-S5		P6-P7 S1-S5		P7 S1-S5
Control w/in district	P4-P7 S1-S5	P4-P7 S1-S5		P5-P7 S1-S5		P6-P7 S1-S5		P7 S1-S5
Control out district	P4-P7 S1-S5	P4-P7 S1-S5		P5-P7 S1-S5		P6-P7 S1-S5		P7 S1-S5
# of schools	234 P 50 S	234 P 50 S		234 P 50 S		234 P 50 S		234 P 50 S
Cohort 2 (8 LANGUAGES)								
Treatment			P4-P7 S1-S5	P4-P7 S1-S5		P5-P7 S1-S5		P6-P7 S1-S5
Control w/in district			P4-P7 S1-S5	P4-P7 S1-S5		P5-P7 S1-S5		P6-P7 S1-S5
Control out district								
# of schools			TBD	TBD		TBD		TBD
Cohort 3 (12 LANGUAGES)								
Treatment					P4-P7 S1-S5	P4-P7 S1-S5		P5-P7 S1-S5
Control w/in district					P4-P7 S1-S5	P4-P7 S1-S5		P5-P7 S1-S5
Control out district								
# of schools								
Total # of schools	384	384	TBD	TBD	TBD	TBD		TBD

E. DATA QUALITY ASSESSMENT

For the USAID/Uganda SHRP P&IE, NORC will take a systematic approach to Data Quality Assessment by ensuring that all systems, protocols, and tools are developed to ensure the highest possible data quality. Data Quality Assessment is therefore a process that includes both formal and informal on-going review of the design of sampling, data collection instruments, field procedures, quality control protocols, and reporting mechanisms that are prerequisites for rigorous external performance and impact evaluation. NORC will document its findings from the DQA in the Semi-Annual reports to USAID.

Following USAID’s DQA guidelines, five key data quality standards will be used to assess quality:

- *Validity*: Do the data clearly and adequately represent the intended result?
- *Reliability*: Do data reflect stable and consistent data collection processes and analysis methods over time?
- *Precision*: Are data sufficiently precise to present a fair picture of performance and enable management decision-making at the appropriate levels?
- *Integrity*: Do the data collected, analyzed and reported have established mechanisms in place to reduce manipulation or simple errors in transcription?
- *Timeliness*: Are data timely enough to influence management decision-making?

In addition to these overarching standards, NORC will review all datasets using three criteria; data must be complete, accurate and internally consistent. NORC’s activities will include reviews of interim datasets, such as data from the post-training pilot, and data from the first week of data collection, so that feedback may be given to the research and data collection teams for mid-course corrections.

Annex 2 includes a list of all data collection instruments and documents that NORC will review as part of its DQA, along with a checklist of DQA items.

F. WORKPLAN

Below we present a detailed work plan schedule for activities for the period of January through September 30, 2013 of Year 1.

Work Plan	
Activities	Month
<p>➤ Deliverable: Performance and Impact Evaluation Design and Year 1 Work Plan</p> <ul style="list-style-type: none"> — Design Report submitted to USAID — Engage in additional communications with implementer, as necessary, based on comments/feedback — Submit revised report 	<p>January 31, 2013</p>
<p>Support sample selection and baseline data collection, including data quality assessments</p> <p>NORC will continue working closely with the SHRP M&E team as they prepare and conduct baseline data collection. Specific NORC activities during this period will include:</p> <ul style="list-style-type: none"> — Conduct or assist the selection of schools for evaluation sample, 	<p>January-March 2013</p> <p>Note: At present, SHRP’s data collection timeline is as follows:</p>

Work Plan	
Activities	Month
<p>ensuring that treatment and control schools are selected according to the sampling plan</p> <ul style="list-style-type: none"> — Review and provide feedback on all data collection instruments – EGRA and learner context instruments, teacher/head teacher survey, classroom observation tool, KAP survey – ensuring that the data being collected link back to evaluation questions/impact indicators, and that the instruments are of high quality — Review and provide feedback on training manuals, data collection protocols/plans, and quality control procedures for field work, and tablet software being used for data collection — Participate, for quality assurance purposes, in enumerator training and pilot testing of instruments – NORC’s Senior Literacy Expert and Survey Specialist will travel to Uganda to participate in the separate trainings for the Results 1 and Results 2 data collections — Conduct field observations during the first two weeks of data collection – NORC’s Survey Specialist, Resident Evaluation Manager, and Senior HIV/AIDS Specialist will travel to the field to observe field work for both Results 1 and Results 2 data collections — Conduct additional field observations in later stages of data collection – to be undertaken by NORC’s Resident Evaluation Manager, and Senior HIV/AIDS Specialist — Conduct quality reviews of data, as it is uploaded onto RTI’s servers from tablets; we expect to be able to conduct real-time data quality reviews if RTI provides NORC with access to the server to which data is being uploaded daily <p>NORC will provide direct feedback to the SHRP implementation and M&E team on any observed issues/concerns; additionally, all quality control activities and findings/observations, and mitigating actions will be documented in the Data Quality Assurance Section of the Semi-Annual Report #1.</p>	<p>Result 1: Training – Feb 11-15 Field work – starts Feb 18</p> <p>Result 2 (KAP Survey): Orientation & training– Feb 25-March 1 Field work – March 4-22</p> <p>NORC experts’ travel will be timed to coincide with these dates</p>
<ul style="list-style-type: none"> ➤ Deliverable: Semi-Annual Report #1, with Data Quality Assessment of baseline data collection ➤ Deliverable: Analysis of baseline data Given the timing of data collection (February through the end of March), it is unlikely that the complete baseline datasets will be ready in time for an analysis to be completed by April 30, as specified in the contract deliverable schedule; a more realistic date for this baseline analysis would be May 30. The analysis of baseline data will be designed, most importantly, to ensure the similarity between treatment and control groups. The analysis will present descriptive statistics for all indicators of interest (impact indicators, covariates) for treatment and control groups, describe any concerns with the data, and suggest adjustments to the impact evaluation design, if necessary. 	<p>April 30, 2013</p> <p>May 30, 2013</p>
<p>Monitor Implementation of Literacy & Health Education Program and</p>	<p>April – Sep 2013</p>

Work Plan	
Activities	Month
<p><i>Evaluation Design</i></p> <p>NORC’s Resident Evaluation Manager and Senior HIV/AIDS Evaluator will make regular visits to intervention schools to observe the implementation of Result 1 and Result 2 activities as a means of gathering information for the performance evaluation, such that the midterm and final performance evaluations will not be informed by information collected only at two points in time, but instead will be fed organically by information gathered in real time, throughout the project.</p> <p>These field visits and regular meetings with SHRP Results Teams will also serve to ensure that program implementation is adhering to the impact evaluation designs. Where implementation deviates from plan, NORC will discuss with USAID and RTI, and discuss adjustments to implementation plan and/or evaluation design. NORC will engage in ongoing dialogue and discussion with RTI about evaluation implementation, and challenges.</p> <p>Semi-Annual and Annual Reports will include details of evaluation implementation, problems encountered, midstream course correction/modification to design and findings during the reporting period.</p>	

ANNEX 1: SAMPLE SIZE CALCULATIONS

Reading Program (Result 1)

The objective is to construct a double-difference estimate of impact, based on a pretest-posttest-comparison-group design. Sample surveys will be conducted at three times: baseline, midterm and endline. The case of sampling for proportions is considered, in which it is assumed that at baseline the treatment and comparison groups each have population proportion values equal to 0.05; at midterm the treatment group has population proportion 0.35 and the comparison group has population proportion 0.18; and at endline the treatment group has population proportion .8 and the comparison group has population proportion 0.40. These values are RTI working assumptions provided by Benjamin Piper and are based on those observed in a similar project in Kenya.

Statistical formulas

The double-difference measure is the difference, between the treated and control populations, of the difference in population means before and after the program intervention. The formula for the power of a one-sided test of the test of the hypothesis that the double-difference measure of impact is zero, when it is in fact equal to D, is:

$$\text{Power} = 1 - \beta = \text{NORMSDIST}[D/\text{sqrt}(\text{varest}) + z_{\alpha}]$$

where

NORMSDIST(.) = normal probability function (i.e., it returns the probability that a normal deviate is less than or equal to the argument)

α = probability of making a Type I error of rejecting the (null) hypothesis when it is true; assumed to be $\alpha = .05$.

β = probability of making a Type II error of accepting the hypothesis when it is false

D = minimum detectable effect

z_{α} = value of normal deviate corresponding to probability α (i.e., z_{α} is the ordinate having probability α to the left); for the values $\alpha=.975$ and $.025$, these values of z_{α} are 1.96 and -1.96, respectively.

varest = variance of impact estimator.

The reason for using a one-sided test of hypothesis is that in most evaluation studies the direction of change is specified (i.e., the direction change in an indicator of interest is desired or expected to be in a specified direction, not either direction). For $\alpha = .05$, the value of z_{α} is -1.6449. For the situation being considered, the value of varest (the variance of the impact estimator) is given by:

$$\text{varest} = [\sigma_1^2 + \sigma_2^2 + \sigma_3^2 + \sigma_4^2 - 2\rho_{12}\sigma_1\sigma_2 - 2\rho_{13}\sigma_1\sigma_3 + 2\rho_{14}\sigma_1\sigma_4 + 2\rho_{23}\sigma_2\sigma_3 - 2\rho_{24}\sigma_2\sigma_4 - 2\rho_{34}\sigma_3\sigma_4]/(n/\text{deff})$$

where

the four design groups are designated by the indices 1 (treatment before), 2 (treatment after), 3 (comparison before) and 4 (comparison after)

n = sample size for each design group (assumed here to be the same for each group)

σ_i^2 = population variance for group i

ρ_{ij} = coefficient of correlation between groups i and j

$deff$ = Kish's design effect (to reflect the effect of multistage sampling ("clustering")) ($deff$ is the ratio of the variance of the estimator under the design to the variance using a simple random sample of the same size).

If the power is specified, the formula for the sample size is obtained by specifying a value for β , e.g., $\beta = 0.1$ (corresponding to a power of $1 - \beta = 0.9$), and solving the formula given above for n .

It is assumed that the value of the intra-school correlation coefficient (icc) is 0.1, and that a sample of $m=30$ P1 students will be selected from each school. (Note that this parameter (icc) reflects the intraunit correlation from *all* levels of sampling.) In this case, the value of $deff$ is $deff = 1 + (m-1)icc = 1 + (30-1)0.1 = 3.9$. It is assumed that there will be no matching of districts or schools, so that ρ_{13} , ρ_{14} , ρ_{23} , and ρ_{24} are zero. The values of ρ_{12} and ρ_{34} are assumed equal to .3, corresponding to a modest degree of temporal correlation (associated with the fact that the follow-up surveys would use the same school samples). A value of $icc = 0.065$ was observed in a similar study ; a slightly larger value ($icc = 0.1$) is assumed here. This assumption is "conservative," since the required sample size increases as the value of icc increases.

Sample size estimates

For the situation described above, the effect size, D , is calculated as the double-difference of the population proportions for the various design groups. Comparing the midterm to the baseline, the double difference is $D = (0.35 - 0.05) - (0.18 - 0.05) = 0.17$. Comparing the endline to the baseline, the double difference is $D = (0.8 - 0.05) - (0.4 - 0.05) = 0.4$. Comparing the endline to the midterm, the double difference is $D = (0.85 - 0.35) - (0.4 - 0.18) = 0.23$.

We shall estimate the sample sizes required to detect the preceding effects with 90% power (probability).

Case 1. Comparing midterm to baseline. The values of the various parameters involved in the power formula are as follows. $\sigma_1 = \sqrt{.05 \cdot .95} = .218$; $\sigma_2 = \sqrt{.35 \cdot .65} = .477$; $\sigma_3 = \sqrt{.05 \cdot .95} = .218$; $\sigma_4 = \sqrt{.18 \cdot .82} = .384$. Assume $\alpha = .05$ and $\beta = .1$ (i.e., power = $1 - \beta = 90\%$). Design effect $deff = 3.9$ (as discussed above). Intergroup correlations (ρ 's) specified as discussed. It is desired to detect a double-difference measure of impact of magnitude $D = .17$. In this case, the required sample size for each of the four design groups -- treated before, control before, treated after, control after-- is 414 pupils. With a sample size of $m=30$ pupils per school, the school sample size is 14 schools. This is the number of schools required for each of the 12 design "cells" (combinations of 4 languages by 3 treatment arm), so that the total number of schools required is $12 \times 14 = 168$.

Case 2. Comparing endline to baseline. The values of the various parameters involved in the power formula are as follows. $\sigma_1 = \sqrt{.05 \cdot .95} = .218$; $\sigma_2 = \sqrt{.8 \cdot .2} = .400$; $\sigma_3 = \sqrt{.05 \cdot .95} = .218$; $\sigma_4 = \sqrt{.4 \cdot .6} = .490$. Assume $\alpha = .05$ and $\beta = .1$ (i.e., power = $1 - \beta = 90\%$). Design effect $deff = 3.9$ (as discussed above). Intergroup correlations (ρ 's) specified as discussed. It is desired to detect a double-

difference measure of impact of magnitude $D = .4$. In this case, the required sample size for each of the four design groups (treated before, control before, treated after, control after) is 80 pupils. With a sample size of $m=30$ pupils per school, the school sample size is 3 schools. This is the number of schools required for each of the 12 design “cells” (combinations of 4 languages by 3 treatment arms), so that the total number of schools required is $12 \times 3 = 36$.

Case 3. Comparing endline to midterm. The values of the various parameters involved in the power formula are as follows. $\sigma_1 = \text{sqrt}(.35 .65) = .477$; $\sigma_2 = \text{sqrt}(.8 .2) = .400$; $\sigma_3 = \text{sqrt}(.18 .82) = .384$; $\sigma_4 = \text{sqrt}(.4 .6) = .490$. Assume $\alpha=.05$ and $\beta=.1$ (i.e., power = $1 - \beta = 90\%$). Design effect $d_{eff} = 3.9$ (as discussed above). Intergroup correlations (ρ 's) specified as discussed. It is desired to detect a double-difference measure of impact of magnitude $D = .23$. In this case, the required sample size for each of the four design groups (treated before, control before, treated after, control after) is 346 pupils. With a sample size of $m=30$ pupils per school, the school sample size is 12 schools. This is the number of schools required for each of the 12 design “cells” (combinations of 4 languages by 3 treatment arm), so that the total number of schools required is $12 \times 12 = 144$.

Conclusion. To detect double-difference measures of the size anticipated, a sample of about 12-14 schools is required, under the assumptions made. This is the sample size for each design “cell” (language by arm combination). A school sample of size about 150 schools (for all 12 design “cells”) should be adequate to detect effects (double-difference impacts) of the anticipated magnitude with high power (90%).

School Health Program (Result 2)

As before the objective is to construct a double-difference estimate of impact, based on a pretest-posttest-comparison-group design. We considered 2 cases using difference assumptions about knowledge, attitudes and practices regarding HIV and AIDS.

Case 1: The case of sampling for proportions is considered, in which it is assumed that at baseline the treatment and comparison groups each have population proportion values equal to 0.7 and at endline the treatment group has population proportion 0.9 and the comparison group has population proportion 0.7.

It is the opinion of Mr. Frank Rewekikomo from RTI that this assumption reflects the HIV and AIDS knowledge proportions and program effects.

Case 2: it is assumed that at baseline baseline the treatment and comparison groups each have population proportion values equal to 0.5 and at endline the treatment group has population proportion 0.6 and the comparison group has population proportion 0.5.

Although it is difficult to predict this second set of assumptions seems more appropriate to evaluate attitudes and practices.

Using the same formulas we described under Result 1 above, power is specified at 0.9, intra school correlation = 0.1 and 30 pupils per school, case 1 assumptions require a sample size of 15 schools. Under the same parameters, the assumptions described in case 2 required a sample size of 78 schools

Conclusion. To detect double-difference measures of the size anticipated for all dimensions - knowledge, attitudes, and practices- a sample of about 78 schools is required, under the assumptions made. This is the sample size for each design arm (treatment, control within district and control outside district) A school sample of size about 234 schools should be adequate to detect effects (double-difference impacts) of the anticipated magnitude with high power (90%).

In addition to 78 primary schools per arm, we will evaluate 50 post-secondary schools where the intervention will also take place. This number of post-secondary schools is the total number of schools that will receive treatment. Additionally 50 post-secondary schools that serve as controls within intervention districts and another 50 will serve as control in control districts.

ANNEX 2: DATA QUALITY ASSESSMENT CHECKLIST

The following data collection instruments, documents and activities will be reviewed as part of NORC’s DQA. Note that this list may not be exhaustive. This list applies to both Result 1 (literacy) and Result 2 (health).

Instrument, Document or Activity	DQA checklist
Evaluation Design	
<i>Sampling:</i> sample size, selection (randomization process) of intervention schools, matching of comparison districts	<ul style="list-style-type: none"> ▪ Sample size is sufficient for desired level of precision and power ▪ Sample design is adequate for assessing impact of intervention at school and district level (randomization of intervention schools is carried out correctly, selection of matched comparison districts is done using adequate statistical matching methods and with best matching data available)
<i>Data collection plan:</i> timing of data collection, selection of intervention/control in-district/control out-district schools for data collection	<ul style="list-style-type: none"> ▪ Timing of data collection is appropriate for impact evaluation (baseline is prior to intervention, follow-ups are at regular intervals, endline is post-intervention; data collections happen at either beginning or end of school year) ▪ Data collection team has allocated adequate human and material resources to carry out collection within specified time period ▪ Timing of data collection and data delivery allow for annual impact evaluation and impact evaluation of 4-year SHRP within project deadlines ▪ IRB permissions have been obtained
Data Collection Instruments	
<i>Result 1 (Literacy)</i> EGRA tool Learner environment questionnaire Teacher/Head teacher surveys Classroom observation tool School survey CCT Monitoring Tool	<ul style="list-style-type: none"> ▪ All tools capture information needed to calculate key indicators for performance and impact evaluation ▪ Questionnaires are ordered logically and structured to facilitate comprehension by respondents and use by data collectors ▪ Questionnaires are piloted and revised accordingly (adapted to Ugandan context) ▪ Questionnaires include proper geo-referencing information and allow for easy merging of data:
<i>Result 2 (Health)</i> KAP survey	<ul style="list-style-type: none"> ✓ Questionnaires include case id, class id school id that are standard across different instruments ✓ Questionnaires are designed for easy merging of longitudinal data ▪ Questionnaires allow capture of interviewer id, supervisor

Instrument, Document or Activity	DQA checklist
	id, data enterer id (if applicable) <ul style="list-style-type: none"> ▪ Observation/review of pre-test results for tool development where applicable ▪ If possible, check that translations have been done correctly (may not be possible due to lack of staff with knowledge of local languages)
Training and Data Collection Period	
Enumerator and Supervisor Training Manuals	<ul style="list-style-type: none"> ▪ Cover at a minimum: project description, basic interviewer techniques, confidentiality and consent, organization of fieldwork and sample requirements, tracking of sample, detailed description of data collection tools
Field Quality Control Procedures	<ul style="list-style-type: none"> ▪ Organization of field teams provides adequate supervision/management ▪ Validation (back-check) procedures are included ▪ Field procedures includes proper tracking of sample and response rates: documentation of in-field sampling procedures (e.g. random selection of students within each classroom), proper use of disposition codes ▪ Interviewer feedback process is documented and used ▪ Proper use of unique ID codes for schools, students, etc, to allow for triangulation of data ▪ Mechanisms for reporting to Central Office/Level of supervision from Central Office is adequate ▪ Schedule for validation, tracking and interviewer feedback reports is clear
Enumerator/Supervisor training for Result 1	<ul style="list-style-type: none"> ▪ Observations of trainings by NORC expert(s) ▪ Trainings are well-organized and trainers are well-prepared
Enumerator/Supervisor training for Result 2	<ul style="list-style-type: none"> ▪ Role-playing and other practice exercises are included ▪ Interviewers demonstrate mastery of concepts and procedures through formal, documented assessment
Post-training pilot	<ul style="list-style-type: none"> ▪ Observation of post-training pilot by NORC expert ▪ Enumerators are well-prepared for field period ▪ Participation of NORC in post-pilot debriefings to gather lessons learned ▪ Debriefing lessons are implemented and communicated to field team prior to field period
Data Collection Field Report	<ul style="list-style-type: none"> ▪ Data collection process and issues encountered during field period are documented (organization and structure of field teams, dates of field report, final response rates,

Instrument, Document or Activity	DQA checklist
	reasons for non-response, challenges encountered and solutions)
Data Entry	
Tangerine data entry template for all applicable tools	<ul style="list-style-type: none"> ▪ Paper instruments are reviewed for completeness prior to data entry ▪ Procedures for handling missing data are clearly specified and standardized across instruments and rounds of data collection ▪ Data entry templates match paper instruments ▪ Skips are respected
Data entry templates for other surveys using paper questionnaires	<ul style="list-style-type: none"> ▪ For paper instruments, data is entered using double data entry method ▪ Soft and hard validation checks are programmed and tested prior to training for electronic instruments, and prior to data entry for paper instruments ▪ Upload are made available for review by data quality reviewer on a real-time basis (as they are uploaded)
Datasets	
EGRA data Learning environment data Teacher/Head teacher survey data Classroom observation data School survey data CCT monitoring data	<ul style="list-style-type: none"> ▪ Datasets are well constructed: variable names, variable labels, value labels are included and correctly specified ▪ Datasets can be easily merged if needed (using of unique codes for merging across different datasets) ▪ Reserve codes are correctly used (including specification of legal skips and missing values) ▪ Proportion of missing values is within acceptable range ▪ Level of precision is adequate ▪ Data is internally consistent <p>(See Annex 1 for more information on Guidelines for Data Cleaning and Assessment)</p> <ul style="list-style-type: none"> ▪ Test datasets and interim datasets (pilot dataset, first 100 cases) are produced and delivered for DQA with adequate time for incorporating corrections prior to main data entry (for paper instruments)
Note: For all instruments, DQA covers pilot datasets, interim datasets (real-time uploads onto Tangerine server), full datasets	
Description of achieved sample sizes, calculation of response rates with breakdown of disposition codes	<ul style="list-style-type: none"> ▪ Realized response rates are adequate to maintain level of precision and power needed for the impact evaluation ▪ Reasons for non-response are well-documented using standard codes ▪ Calculations of sample weights, when needed, are done

Instrument, Document or Activity	DQA checklist
	correctly
Data documentation	<ul style="list-style-type: none"> ▪ Proper metadata (see Annex 2) is included with the datasets (codebook at a minimum) ▪ Documentation of any cleaning steps taken before delivery of final datasets ▪ Delivery of both raw and cleaned datasets, deidentified, if required