

DRAFT – ACR ROUND 1 EVALUATION ASSESSMENT AND FINDINGS
USAID Reading and Access Evaluation Contract
NORC at University of Chicago

I. Introduction and methodology

Under the *All Children Reading: A Grand Challenge for Development* Round 1 (ACR1) grant, grantees reported on a number of output and outcome indicators for Monitoring & Evaluation purposes. One of the ultimate outcomes of interest across all grants is whether beneficiaries' reading skills improved through the course of the intervention. Funders, implementers, researchers, and other stakeholders are interested in knowing:

- Did interventions under the All Children Reading Round 1 grants report improved reading outcomes?
- To what extent are the reporting reading outcomes validated by a rigorous evaluation?
- Can these reported reading outcomes be generalized to the target population?

I.A Assessing the validity of findings

Our ability to draw conclusions about the above two questions depends largely on the design of the evaluation, and the associated sampling, data collection, and analysis. In order to be able to affirmatively answer these two questions, impact evaluations need to meet two important criteria. They must have a high degree of:

- **Internal validity:** A requirement for being able to attribute changes in reading outcomes to an intervention. If we observe an increase in reading outcomes, and have examined and isolated other causes of this increase, then we can attribute part of the increase in reading outcomes to the intervention. Common threats to internal validity among the Round 1 grantee evaluations include the lack of a proper counterfactual, sample selection bias, attrition, contamination, and lack of data or analysis on whether and how other observable or unobservable characteristics might have influenced reading outcomes.
- **External validity:** A necessary component of being able to generalize results of a study to a target population is assessing whether the sample of the study is representative of the population that would be targeted by any scale up of the intervention. Common threats to external validity among Round 1 grantee evaluations include: non-representative sample and implementation that did not occur as originally planned. Note that a single study, no matter how rigorous, can rarely be generalized to a broader population. However, some studies make claims that the results observed could be replicated on a larger scale, when in actuality, the design and sampling of the evaluation only allows the study to make claims about a very narrow subset of the population. Understanding threats to external validity helps evaluate the degree to which these interventions can be expanded or scaled up.

Grantees addressed the two research questions using several different methodologies and approaches. However, not all of the resulting approaches were internally and externally valid. To determine whether an evaluation was internally or externally valid, NORC researchers reviewed information from endline evaluation reports, baseline evaluation reports, sampling design

DRAFT – ACR ROUND 1 EVALUATION ASSESSMENT AND FINDINGS
USAID Reading and Access Evaluation Contract
NORC at University of Chicago

documents, write-ups on the data collection instruments, and reporting on implementation fidelity for the following information:

- **Evaluation design.** Generally, evaluation designs that reduce threats to internal validity included experimental designs (randomized controlled trials), and quasi-experimental approaches that require a number of credible assumptions to be met. For example, propensity score matching assumes that if observable characteristics of treatment and control groups are very similar, the same would apply to unobserved characteristics. Other quasi-experimental approaches require stronger assumptions to be true. Difference-in-difference studies, for example, assume that in the absence of the intervention, treatment and comparison scores, even if different at baseline, would move in tandem; as a result, claims of causality need to be interpreted with caution unless auxiliary convincing evidence is provided. Weaker designs, such as pre/post tests without a comparison group are unable to demonstrate a causal relationship between the intervention and reading outcomes.
- **Sampling methodology.** Sampling strategies and sample sizes were examined closely to determine whether the sample seemed large enough (we did not have all the information needed to do rigorous sample size calculations) and free of selection issues. In general, studies that hand-picked a subset of schools because they were underserved and/or had adequate infrastructure, or recruited subject participants on a first-come, first-serve basis, present selection problems. In RCT studies, we also focused on baseline differences between treatment and control groups to check whether the sample was properly balanced. For difference-in-difference studies, we considered the likelihood of time variant characteristics differing between treatment and comparison groups, or whether there were any mitigating factors taken into consideration when selecting a comparison group to better account for these differences.
- **Implementation fidelity/quality.** If outcomes did not meet expectations, it could have been that the intervention did not work or that it was not implemented as originally intended. We examined data on implementation fidelity to help examine why outcomes may have differed from what was hypothesized, and if implementation issues were substantial enough that this study may not be replicable to other contexts.
- **Data Collection.** Because we were ultimately interested in reading outcomes, we examined the instruments used to assess these outcomes– including the Early Grade Reading Assessment. We examined whether the instrument used for data collection was tested, whether the instrument changed between data collection rounds, thereby making it less comparable between baseline and endline (or, if the instrument did not change, whether teachers may have been teaching to the test or students remembered the test from the previous round of data collection), and whether there were unexpected events in data collection that led to systematic missing data.
- **Analysis.** Inferential testing (e.g. tests for statistical significance) helps determine whether or not results are due to chance, or will likely happen again. Analyses that included

DRAFT – ACR ROUND 1 EVALUATION ASSESSMENT AND FINDINGS
USAID Reading and Access Evaluation Contract
NORC at University of Chicago

inferential testing and accounted for observable characteristics that can influence reading outcomes outside of the intervention had higher levels of internal validity than those that did not.

I.B Methodology

To systematically review each evaluation report and assess the validity of each evaluation’s findings, a team of practitioners and economists from NORC with experience designing and executing impact evaluations reviewed each evaluation using the Evaluation Assessment Framework (Annex 1). To ensure that the use of the framework in assessing the evaluation was consistent, the team piloted the framework with several assessors rating the same study and comparing results to test inter-rater reliability. Where information was unavailable in an evaluation report, the team reviewed baseline reports, instruments, and other documents and noted if the information was available. Based on the review of each evaluation report and supporting documents, we provide an assessment of the degree of caution that should be exercised in interpreting the findings of each report.

Note that the team only read reports from 14 out of the 32 ACRI grantees. Grantees were excluded from this assessment if they did not have an endline assessment, will not have an endline report completed before the publication of this report, or did not have a report with data that was conducive for this assessment.

I.C Missing information

For this assessment, we relied solely on the information available in the documentation provided by grantees. Due to time constraints, we were unable to follow up with grantees to obtain missing information as it relates to the evaluation, or further clarify statements that were confusing or unclear. When reviewing the assessment, it is important to note that if a specific item of information required for this assessment was not available in a report, it did not necessarily mean that the evaluation did not account for that information. For example, many reports did not include detailed information on how districts, schools, and students were selected for the sample. Other reports did indicate that the selection of respondents was not random (e.g. first-come, first-serve basis), which poses a clear threat to the validity of the findings. Where possible, we have tried to make a distinction between a piece of information that leads us to believe we should proceed with caution, or whether we need to proceed with caution because we lack a certain piece of information. If information critical to drawing a conclusion about the intervention was missing, that was noted in the analysis. See the “Threats to Validity” column in Table 1 for a more nuanced discussion of what details of the evaluation were missing from reports.

I.D What this assessment does not cover

Many grantee reports were comprehensive in outlining implementation fidelity, as well as collecting and reporting on intermediary steps and outcomes in their theory of change. These

DRAFT – ACR ROUND 1 EVALUATION ASSESSMENT AND FINDINGS
USAID Reading and Access Evaluation Contract
NORC at University of Chicago

crucial pieces of analysis not only help better structure future programs, but also help us understand why results did or did not meet expectations. However, this report did not assess evaluations from the standpoint of intermediary outcomes, such as parent and community engagement/behavior change, and teacher uptake; it focused solely on reading outcomes.

Furthermore, this assessment does not comment on the quality of the reports themselves, including data visualization, logical structuring of the reports, and completeness of the reports. Assessors focused solely on reviewing available information to the best of their ability to determine whether there were threats to the internal and external validity of the evaluations. However, there were times where the quality of the report hindered our ability to fully understand the methodology and analysis—where this was an issue, it has been noted.

Finally, this assessment is not a meta-analysis—that is, it does not compare effect sizes across ACR1 grants, and is not meant to analyze the reported results to integrate the findings and make broad statements about the effectiveness of the ACR1 grant initiative. Instead, it is focused on evaluating the results claimed made by each grantee, and understanding what conclusions we can and cannot arrive to about the effects of each grant based on the available evidence.

II. Assessment of evaluation findings

Below, we present a discussion of the threats to internal and external validity by each grantee evaluation, along with an explanation of whether the increased outcomes can be attributed to the intervention, and whether or not the results could be generalizable to the target population. In Table 1, we summarize our degree of confidence in the findings presented in each report. Table 1 provides a brief overview by each grantee about the findings reported, the type of design, our assessment of whether the grantee reported an increase in outcomes, whether a claim about the causal relationship between the intervention and the outcome is valid, and whether the sampling of the study made it conducive to generalizing results to a broader population. For a more detailed overview of each grant, the elements of the intervention, beneficiaries, evaluation design, sampling, data collection and implementation quality, and validity of conclusions side-by-side, see Table II in the Appendix. <<See attached excel file, to be inserted into report. Appears as Table 1 in the Internal Report>>.

II.A Overall Quality of Evaluations

Of the 14 studies we reviewed, we have information about the evaluation design of 13; four are RCTs, four are quasi-experimental, difference-in-difference studies, and five are one-group, pre-/post-test. Generally, well-executed randomized controlled trials and quasi-experimental designs that have a clear counterfactual are likely to have higher degrees of validity than a pre/post-test without a comparison group. However, the four quasi-experimental studies assessed do not provide adequate justification for us to determine if the assumption that treatment and comparison scores would move in tandem in absence of the intervention is valid. Of the four evaluations that employ RCTs, only one tests for balance between treatment and control groups at

DRAFT – ACR ROUND 1 EVALUATION ASSESSMENT AND FINDINGS
USAID Reading and Access Evaluation Contract
NORC at University of Chicago

baseline. Given that samples are not very large, balance with the treatment group is an important characteristic of a proper counterfactual. Across nine of the 13 reports, selection bias issues hinder our ability to generalize the study to the target population.

Based on our assessment of the available information, it is our conclusion that most of the ACRI grant evaluations did not have high degrees of internal and external validity. However, the design and execution of some studies were more rigorous than others; to make it easier to navigate the different results from each study, we've grouped studies by the evaluation design employed.



II.B What, if anything, can we conclude from these reports?

As discussed above, the criteria for understanding the impact of ACRI reports stems from the degree of internal and external validity of the findings- with what degree of confidence can we say that the intervention led to increased outcomes, and that these results could be generalized to a larger target population? As Table 2 demonstrates, in most cases, reading outcomes for the recipients of the intervention increased over time. However, due to threats to internal validity, there is no overwhelming evidence that the ACRI interventions we reviewed were responsible for this increase in reading outcomes. In fact, we cannot make sound judgments about attribution and causality of almost all of the ACRI interventions based on the available information; more rigorous research is needed to determine whether and how these interventions increase reading outcomes. So while we can say that, on average, reading outcomes increased for recipients of ACRI grant interventions, how much of the increase is attributable to the interventions cannot be determined from the available data. Furthermore, we do not know whether these results are localized to the sample, or whether they can be generalized to the target population.





DRAFT – ACR ROUND 1 EVALUATION ASSESSMENT AND FINDINGS
 USAID Reading and Access Evaluation Contract
 NORC at University of Chicago

Table 2: Validity of Findings Across ACR1 Grantee Reports



Refer to the Grantee Annexes for a more detailed description of the summary of the intervention and other detailed information about activities

Grant	Reported Findings	Design	Can we conclude that the intervention led to increased outcomes?
Urban Planet Uganda: MobiLiteracy	<p>For students in Primary 1 and Primary 2:</p> <ul style="list-style-type: none"> - Treatment arms had higher, statistically significant increases in scores over control in familiar word reading (mobile: 55% increase higher than control, paper: 69% increase higher than control) and listening comprehension (SMS: 29% increase higher than control, paper: 48% increase higher than control). -However, gains are incremental: for example, in familiar word reading, scores increased by 1.4 out of 10 points for the mobile group, and 1.5 out of 10 points for the paper group - Endline gains also higher but not statistically significant for the paper-arm versus the mobile-arm relative to the control (letter sound identification, mobile: 18%, paper: 28%, nonword decoding, mobile: 0%, paper: 42%). Both arms did worse than control on syllable segmentation. - Effects decrease or perform worse than control when observing children that have some reading ability (scored greater than 0 on an EGRA subtask); none of the increases are statistically significant -Reduction in the percent of zero scores in letter sound identification (mobile: -44%, paper: -53%) and listening comprehension (mobile: -66%, paper: -78%) from baseline to endline relative to control is higher for mobile and paper-based groups (with paper outperforming mobile), relative to the control 	RCT	
			<ul style="list-style-type: none"> -Well-executed randomized controlled trial -Only two gains are statistically significant: familiar word reading and listening comprehension -Difference-in-difference regression analysis would have provided more precision
Worldreader Ghana: iREAD	<p>For students in Primary 1 through Primary 3:</p> <ul style="list-style-type: none"> - Statistically significant difference in the gain in correct words per minute (cwpm) for treatment students, compared to control -Twi reading comprehension: treatment students improved 23 pp compared to the control group's improvement of 13 pp 	RCT	
			-No balance tests performed





DRAFT – ACR ROUND 1 EVALUATION ASSESSMENT AND FINDINGS
 USAID Reading and Access Evaluation Contract
 NORC at University of Chicago

	- English reading comprehension: treatment students improved 27 pp compared to the control group's 23pp			
Ghana Reads Project, Ghana (Open Learning Exchange)	For students in K through Primary 3: -Greater increase in scores for treatment versus control schools across all subtasks. For example: -Oral test (score, number correct) treatment increases from 6.41 to 41.69, control increases from 10.69 to 21.19 -Reading comprehension (number correct): treatment increases from .36 to 2.77, control increase from .5 to 1.3 -Overall score: treatment increase from 35.6 to 177.8, control increases from 40.1 to 92.7	RCT		
			-No information on whether groups balanced at baseline. Not enough information to properly assess its validity	
All Children Reading, Somalia (African Education Trust)	For students in Grades 1 and 2: - Pilot schools: 58% of students were unable to move past section 3 (the transition from picture-word recognition to reading simple words and short paragraphs) at baseline; this dropped to 27% at endline -Control schools: at baseline 66% were unable to move past section 3, and at endline 34% were unable to move past-control outperforms by one percentage point	Unclear-likely RCT		
			-No comparison of baseline and endline groups, given that the treatment group had a different set of students tested at baseline and endline and the control group had a different set of students tested at baseline and endline -No inferential testing -Note: contamination of control schools, with teachers and government officials from pilot schools sharing best practices may imply that control school scores are higher than what they would have been had contamination not occurred	
All Children Reading, Rwanda (Drakkar)	For students in Primary 3: - Number of words read: treatment average improved by .092 standard deviations (a 1.28 word/minute increase) over control for number of words read - Listening comprehension: treatment improved by .137 standard deviations greater than control (.2 correct answer more)	DiD		
			-No information provided on pre-intervention trends between treatment and comparison. Strong assumptions required	
All Children Reading, Sri Lanka (Save the Children)	For students in Primary 1 through Primary 3: -Only reading measure that the Literacy Boost intervention positively impacted was "fluency" (0.35 effect size). No other significant difference	DiD		



DRAFT – ACR ROUND 1 EVALUATION ASSESSMENT AND FINDINGS
 USAID Reading and Access Evaluation Contract
 NORC at University of Chicago

	<p>was found between Literacy Boost and comparison groups on any other literacy outcome measure</p> <ul style="list-style-type: none"> -Effect on fluency was stronger for Literacy Boost “slow learners” and slow learners with disabilities than for total sample combined (0.53 effect size) 		<ul style="list-style-type: none"> -No information provided on pre-intervention trends between treatment and comparison. Strong assumptions required -Potential purposive sample of slow-learners -Sampling strategy unclear
<p>Enlightening the Hearts Literacy Campaign Training for Transformation, Ghana (Olinga Foundation)</p>	<p>For students in Primary 4 and Primary 5: - In year 1, percent of students literate at program schools went from 19% to 59%; percent of students literate at non program schools went from 21% to 49%- a difference of 12 percentage points</p>	<p>DiD</p>	
			<ul style="list-style-type: none"> -Systematic missing data for many year 1 students who began program in P6 -Teachers may have volunteered for trainings -Four districts lost all non-beneficiary schools in year 2 -No information provided on pre-intervention trends between treatment and comparison. Strong assumptions required
<p>WhizKids Workshop, Ethiopia (WhizKids)</p>	<p>For students in Primary 2: - Scores seemed to have increased between baseline and midline within the target (treatment) and control groups. -Target schools outperform control schools by approximately 2 points across all subtasks -Letter name fluency: control school scores increased from 51.7 to 69.3 letter names per minute (17.6 point increase), for target schools, scores increased from 51.3 to 70.9 (19.6 point increase) -Familiar word fluency: control school scores increased from 26.5 to 39.2 familiar words per minute (12.7 point increase), target school scores increased from 25.5 to 41 (15.5 point increase) -Invented word fluency: control school scores increased from 15.7 to 21.4 invented words per minute (5.7 point increase), and increased from 15.3 to 22.6 for target schools (7.3 point increase) -Passage reading fluency: control school scores increased from 23.3 to 34 context words per minute (10.7 point increase), while target school scores increased from 22.9 to 35.4 (12.5 point increase)</p>	<p>Unclear–likely DiD</p>	
			<ul style="list-style-type: none"> -Selection method of schools and students unclear -Unclear how proficiency levels were defined -No inferential testing reported -Difficult to assess without knowing evaluation design

DRAFT – ACR ROUND 1 EVALUATION ASSESSMENT AND FINDINGS
 USAID Reading and Access Evaluation Contract
 NORC at University of Chicago

<p>Periodic Learning Camps for Reading Improvement, India (Pratham)</p>	<p>For students in Grades 3-7: -Reading fluency scores for children assessed with Grade 2 level text increased by double-digit percentage point (pp) increases (range of 37 to 74pp) for grades 3-7. For example, reading fluency scores for Grade 3 children went from 9% to 83%, for Grade 7 children scores went from 61% to 98%. -Similar gains in reading comprehension scores for children assessed at PISA level 1 and level 2 (e.g. range of increases 38 to 59 pp for grades 3-4 across subtasks, range 32 to 58pp for grades 5-7)</p>	<p>One group pre/post test (no control)</p>	<p style="text-align: center;"></p> <p>-No control group -No inferential testing -Attrition; no data on how these students performed at baseline relative to rest of sample</p>
<p>Timawerenga! We Can Read, Malawi (FHI360)</p>	<p>For students in Standards 1 and 2: -Text-based reading skills and listening comprehension: only notable improvements in scores -Standard 2 students scored 4.7 cwpm on letter sound identification, 7 correct syllables per minute (cspm) for syllables, and 4.6 cwpm for oral reading fluency—with an increase of 3 to 6 items per minute across all text-based reading skills -Standard 1 students scored less than 1 item per minute at endline across all categories -Reading comprehension scores: not broken out by standard, but at endline only 6% of students were able to answer one or more comprehension questions correctly.</p>	<p>One group pre/post test (no control)</p>	<p style="text-align: center;"></p> <p>-No control group</p>
<p>Let the Children Read in Own Language, Bangladesh (ECo-Dev)</p>	<p>For Class II-Class IV students: -Words read per minute: increase from 16 at baseline to 42 at endline -Percent of children who can read 35-90 words per minute: increased from .75% at baseline to 51% at endline -Percent of students who answered 80% of comprehension questions: 0% at baseline to 40% at endline</p>	<p>One group pre/post test (no control)</p>	<p style="text-align: center;"></p> <p>-No control group -No inferential testing -Indicators are somewhat imprecise, could lead to ambiguity</p>
<p>Development of Bilingual Literacy in Minority Schools, Georgia (DBL)</p>	<p>For students Primary 1 through Primary 6: -Vocabulary, phonics, reading comprehension: Proportion of students in the “low” category decreased from baseline to endline (60.5% to 42.5% for vocabulary, 53.9% to 35.3% for phonics, 64.2% to 60.8% for reading comprehension)</p>	<p>One group pre/post test (no control)</p>	<p style="text-align: center;"></p> <p>-No control group -Purposive sample; sampling potentially not random</p>

DRAFT – ACR ROUND 1 EVALUATION ASSESSMENT AND FINDINGS
 USAID Reading and Access Evaluation Contract
 NORC at University of Chicago

	-This is not the case for grammar (proportion of students in low category increased from 41.4% to 69.5%)		
TiANA, Malawi (Save the Children)	For students in Standard 3: - Baseline scores generally started much lower for English than Chichewa, and children had greater improvements in Chichewa than they did English -Letter sound fluency: 20 pp increase for % correct in Chichewa -Writing: Chichewa score increased 6.3 to 12.4, English increased 1.9 to 3.9 (nominal scores) -Fluency: Chichewa increased from 9.7 to 14.9 wcpm, English increased from 3.5 to 5.7 wcpm. -Reading accuracy, Chichewa % correct scores increased from 31.9 to 46.8; English % correct scores increased from 5.8 to 19.3 -Reading comprehension: Chichewa % correct increased from 22.1 to 46.7%; English % correct from 3.6-9.9%	One group pre/post test (no control)	
			-No control group -Only looks at changes between year 1 and year 2, not the overall program
Same Language Subtitling on TV: India (Planet Read)²	No results reported yet	N/A	
			- Control group is selected from another state -Sample selection, including why some families have channel programming, students selecting into different rates of exposure



Strong assumptions to come to this conclusion; interpret results with caution.



Cannot say “yes” based off information presented.

1 In this question, we are not asking whether the results are generalizable—it is rare that a single study, no matter how rigorous, can be generalized to a broader population. It would need to be replicated in other situations, have a strong theoretical frame, etc. The purpose of this question is to determine whether the sample is representative of the target population—a necessary but not sufficient condition for replication and generalizability.

2 Endline report not yet released; second and third columns are based on evaluation designed, as outlined by the baseline and midline reports.

3 Icons created by Amit Jakhu. Retrieved from <http://www.flaticon.com>. Flaticon is licensed under [Creative Commons BY 3.0](https://creativecommons.org/licenses/by/3.0/). Color of icons altered for purposes of the report

DRAFT – ACR ROUND 1 EVALUATION ASSESSMENT AND FINDINGS
USAID Reading and Access Evaluation Contract
NORC at University of Chicago

II.C What stops us from saying more about ACRI grants from these reports?

Major characteristics of the studies that limit our ability to draw conclusions about ACRI—especially about what types of interventions worked better than others, include the following:

The lack of a clear counterfactual inhibits the ability to attribute results to the intervention in many of these evaluations. When a study only looked at pre- and post-test outcomes for the treatment group, even though we saw increases in reading outcomes, it is not possible to isolate the increases resulting from the intervention versus those due to other factors that are not associated with the intervention (attending school, learning at home). Of the 13 studies that had information about evaluation design, five relied on a pre- and post- test of a treatment group. Of the eight studies that did include a counterfactual (because they were either an RCT or a difference-in-difference), only one adequately demonstrated that treatment and counterfactual are statistically identical at baseline. All these studies assumed that differences in increases in outcomes between treatment and control group were attributable to the intervention. However, as discussed above, several conditions must be met for this to be true. In the case of a difference-in-difference, the assumption is that the comparison group and treatment group would have similar increases in outcome scores in absence of the intervention. None of the reports argue or provide evidence that such an assumption could be made, including the examination of pre-intervention trends. Therefore, we are technically unable to attribute the gain in outcomes of the treatment relative to the comparison group to the intervention. Even in the case of the RCTs, where a counterfactual exists, studies had insufficient sample sizes to adhere to the law of large numbers.

Methods of sampling and selection limit the extent to which findings can be extended beyond the sample to the target population. It also hinders our ability to make a causal claim, especially if the selection criteria for including a student or school in an intervention might be an important factor in determining reading outcomes (for example, low or high performing students), and a similar set of criteria are not used for selecting the control group. In eight out of the 14 studies, insufficient information was provided about the selection process to allow us to make a sound judgment about the quality of the sample. Of the studies that did report information about the selection process, there were sample selection issues. In these cases, because we do not have full information on whether or not the treatment groups in each study are different to the control due to sample selection issues, we cannot make causal claims about the intervention. Two common selection issues that arose were:

- **Initial selection of where to pilot.** Some studies had specific selection criteria about which schools would receive an intervention. However, some of these selection criteria may be correlated with inherent characteristics about districts or schools that make them more conducive to gaining from the intervention. For example, some treatment and control schools were limited to specific geographic locations, districts with a certain level of available resources or a minimum level of infrastructure, or districts that for whatever reason had not received literacy interventions before. Furthermore, some districts were

DRAFT – ACR ROUND 1 EVALUATION ASSESSMENT AND FINDINGS
USAID Reading and Access Evaluation Contract
NORC at University of Chicago

selected by the Ministry of Education or by the implementer without clearly defined criteria. From a practical standpoint, these selection criteria make sense; however, in order to be able to extend the results of an evaluation beyond this restricted sample, the accompanying analysis needs to account for observable characteristics that could influence reading outcomes through regression or other analysis to better understand whether and how much improvements in outcomes can be attributed to the intervention, as opposed to the selection criteria, or the subsequent conclusions need to be caveated/restricted to the appropriate subset of the population.

- **Sample selection of students:** Generalization requires a sample that is representative of the population—samples that include volunteer parents and students are not representative of the population because those that volunteer are different from those that do not in ways that can affect also the outcomes of interest.

Contextualizing results gives more meaning to each finding—without a reference point, it is difficult to discern whether and how much an improvement matters. For example, some studies reported out increases in terms of percentage points; others just provided the average increase in scores (e.g. questions answered correctly, correct words per minute). It is difficult to discern whether and what these increases mean—if the treatment group improved by four more answers correct than the control group, how does that gain fare relative to similar interventions that have been implemented in the country? How much closer does that gain put students to where they need to be performing for that grade level? A few grantees referenced the project goals and metrics set at baseline, and measured the reported gains against these goals, while others cited a similar study in the country and compared the increase experienced in their own grant to those reported by study. These are important starting points for helping contextualize results for the reader. Without these reference points, it was difficult to draw conclusions about how each grant performed.

Finally, this assessment only covers a subset of all ACRI grants. This report only looks at 14 out of the 32 ACRI grantees. This subset of grantee reports is not necessarily representative of all grantees. Other grantees were not included if they did not have an endline assessment, will not have an endline report completed before the publication of this report, or did not have a report with data that was conducive for this assessment.

III. Lessons learned about evaluation quality

The bar by which a causal claim is considered valid and generalizable is high. Often times, it is not possible to have the “perfect” evaluation, but an evaluation of a pilot can still be useful. Furthermore, studies may not necessarily be commissioned to determine whether the intervention led to increased outcomes—or may not aim to generalize the results to a broader population. Clearly laying out research questions, designing a study that is conducive to answering those questions, and being transparent about the claims that can and cannot be substantiated by the

DRAFT – ACR ROUND 1 EVALUATION ASSESSMENT AND FINDINGS
USAID Reading and Access Evaluation Contract
NORC at University of Chicago

study are important in minimizing miscommunication or incorrect assumptions about the study's conclusions. Below are some considerations for grantees as they continue to conduct evaluations:

Is my evaluation design conducive to answering the research questions I've set out to answer? As this report outlines, if the goal is to answer questions about whether the intervention lead to an outcome and by how much, in a pilot evaluation there must be a clear counterfactual. Outline why the chosen design, as well as the chosen analysis, are adequate in answering the research question.

In my report, have I fully explained the following? Help your reader assess the information you have presented by including the following elements, which some ACR1 grantees omitted, to your future reports:

- Research questions
- Evaluation design, and an explanation of why it was chosen
- How treatment and control groups were selected
- How sample size was chosen, including power calculations or other considerations for deciding on the size
- Sampling methodology, including sampling stages, selection method (randomization, matching, etc.)
- Analysis of the balance of the sample when appropriate
- Analysis or discussion of the assumptions required by quasi-experimental methods to be valid
- Any selection bias issues (first come - first serve, volunteer basis, etc.)
- Data collection instrument: whether and how the instrument varied, type of data collection, whether the instrument has been tested and used before, and quality issues encountered during collection
- An assessment of implementation quality, and whether any implementation issues could alter the validity of findings
- Inferential tests, especially when comparing means from treatment and control groups
- Limitations of the study
- Conclusions that are substantiated by the results

Have I made the case that I have a proper counterfactual? If the evaluation is a randomized controlled trial, include information about the balance of the sample at baseline to make a compelling case that the treatment and control groups are statistically identical. If conducting a difference-in-difference analysis, provide evidence or an explanation for why we can assume that any differences in trends between the treatment and comparison groups after the program started can actually be attributed to the intervention and not to other underlying sources of change.

Am I accounting for observable characteristics or events other than the intervention that can impact outcomes? The study should not only account for the sex and age of students, but also collect data on other characteristics that may vary by student, school, or by district (e.g.

DRAFT – ACR ROUND 1 EVALUATION ASSESSMENT AND FINDINGS
USAID Reading and Access Evaluation Contract
NORC at University of Chicago

socio-economic status). In addition, data on implementation quality should also be collected. The analysis needs to take into account other factors that can bring about increased outcomes to determine whether and how much of the changes in outcomes are attributable to the intervention.

Is my sample representative of the target population I hope will receive this treatment? If the intention is to pilot this intervention and later scale up, the sample of districts, schools, and participants in the pilot should have similar characteristics to the population that will receive the scaled up intervention. They should be randomly selected from this population. If that is not possible, outline caveats: are there clear reasons to assume or doubt the results would be similar in other areas (e.g. because the study was conducted with a sample much poorer than others in the target population, is more urban, etc.)

Have I provided context to my results? As discussed earlier, provide a reference point that helps others understand what the significance of the result is. For example, cite similar studies and provide results in those studies as a basis for comparison, or indicate how the gains place students relative to where they should be performing at their grade level.

Am I being open and transparent about the quality, and limitations of my study? Be honest about the limitations of your study; share inconsistencies, limitations, and alternative interpretations. Explain what can and cannot be concluded based off of these limitations.

Am I making a claim about causality, or generalizability, which is unsubstantiated? Unexpected events can occur during the course of an evaluation. It is important to be upfront about what can and cannot be said about the intervention based off of the way the evaluation was designed, and the way that the evaluation played out.

DRAFT – ACR ROUND 1 EVALUATION ASSESSMENT AND FINDINGS
 USAID Reading and Access Evaluation Contract
 NORC at University of Chicago

IV. Annex 1: Evaluation Assessment Framework

1. Study Name:
2. Country:
3. Was the evaluation conducted by an independent evaluator?
4. Who conducted the data collection? Was data collection independent from implementing agency?
5. Project description: <ul style="list-style-type: none"> a. Project goals/objectives: b. Summary of the intervention(s): c. Description of beneficiaries: d. Treatment arms:
6. Evaluation description: <ul style="list-style-type: none"> a. Evaluation objectives (specific to reading outcomes) : b. Evaluation questions/hypotheses (specific to reading outcomes) (fill out in table 2 below) c. Outcome indicators (test scores such as EGRA) (fill out in table 3 below)? d. Evaluation approach / method (RCT, Matching, RDD, before-after comparison, etc.; How were control units, if they exist, selected?)
7. Type of evaluation: <ul style="list-style-type: none"> a. Experimental impact evaluation (RCT) b. Quasi experimental IE with credible assumptions c. Quasi experimental IE with strong (require more conditions to be true) assumptions d. Non experimental quantitative (no counterfactual, before-after outcomes) e. Performance Evaluations (no baseline, implementation/process evaluation)
8. Sampling: <ul style="list-style-type: none"> a. How many sampling stages were used in the data collection? Describe them. b. What is the sample size – total, treatment group, control group? If there are multiple treatment arms, indicate sample size for each arm. c. How was sample size determined? Did the evaluator carry out power calculations? If so, indicate MDES and power used. d. How was the sample selected? Describe the selection method used for primary and secondary sampling units (randomization, matching, panel/refresh, other) e. Analysis of the balance of the sample

DRAFT – ACR ROUND 1 EVALUATION ASSESSMENT AND FINDINGS
USAID Reading and Access Evaluation Contract
NORC at University of Chicago

<p>9. Are there selection problems? If yes, what type of selection problem? <i>i.e. self selection, selection based on some characteristic, first come first served, etc.</i></p>
<p>10. Describe the primary data used for the evaluation:</p> <ul style="list-style-type: none">a. Frequency of data collection (baseline, midline, endline, and dates)b. Type of data collection (e.g. self-reported)c. Describe the instrument, and the quality of the instrument:d. Unit of data collection/analysise. Changes between rounds of collection, if anyf. Enumerator training information if any (including IRR, field-based training/piloting); quality issues encountered if anyg. Quality issues encountered during collection
<p>11. Describe any secondary data used for the evaluation:</p> <ul style="list-style-type: none">a. Frequency of data collection (baseline, midline, endline, and dates)b. Type of data collection (e.g. self-reported)c. Unit of data collection/analysisd. Changes between rounds of collection, if anye. Quality issues encountered during collection
<p>12. Was there any assessment of the quality of implementation? Were there any implementation issues that could alter the validity of findings?</p>
<p>13. Indicate any additional information needed to answer the above questions/require reaching back out to the implementer/evaluator (prioritize in order of importance).</p>
<p>14. Summarize any other strengths, weaknesses, limitations, or concerns (not listed above) with:</p> <ul style="list-style-type: none">- Evaluation design- Sample size estimations (including power calculations)- Sample selection (including selection biases, if any)- Data collection (training, instruments, quality issues, timing)- Implementation quality- Evaluation results- Other comments
<p>15. Validity</p> <ul style="list-style-type: none">a. Were there any threats to internal validity?b. Were there any threats to external validity?
<p>16. Assessment of evaluation quality</p>
<p>17. What are the valid findings and conclusions we can take away from this study?</p>

DRAFT – ACR ROUND 1 EVALUATION ASSESSMENT AND FINDINGS
USAID Reading and Access Evaluation Contract
NORC at University of Chicago

--

Table 2: Research Questions (related to reading outcomes)

Research Question	Methodology and Analysis	Findings	Quality Concerns/Limitations

Table 3: Reading Outcomes

Outcome Measure	Data Collection Method	Quality Concerns/Limitations