2025 LEAP Challenge

LEAP

Project Host:

Fundación Aprender a Quererte

Fundación Aprender a Quererte

Fellows:

Daniel Ansari, Research Fellow

Lissett Babaian, Team Lead, Social Entrepreneur Fellow

Dietsje Jones, Research Fellow

Simi Lawoyin, Social Entrepreneur Fellow

TABLE OF CONTENTS

Executive Summary

Introduction

Fundación Aprender a Quererte (FAAQ) is a new philanthropic platform dedicated to enhancing educational opportunities and fostering economic mobility in Colombia. Over the next few years, FAAQ aims to become an incubator for Colombia's public education sector—identifying, adapting, and rigorously testing educational solutions before partnering with the government to achieve scale.

Organisation's role & strength

FAAQ's unique value proposition lies in its ability to rapidly test and refine education solutions, ensuring they are relevant to local needs, feasible to implement and scale, and effective in helping kids learn.

FAAQ's flagship program, Enseñar al Nivel Adecuado (ENAd), is an accelerated learning initiative aimed at ensuring that all students master foundational literacy and numeracy skills before they transition out of primary school. This program brings Pratham's Teaching at the Right Level (TaRL) approach—successfully implemented across Africa and Asia—to Colombia and Latin America *for the first time.*

Need summary

Through the ENAd program, FAAQ has reached 3,500 children in Bogotá and 4,800 in Risaralda. To monitor, evaluate, and learn from these pilots, FAAQ has adapted several of Pratham's research instruments and developed additional tools to provide a more holistic view. This data has allowed FAAQ to track student and teacher progress and respond quickly when needed. At the same time, FAAQ's current MEL system revealed several gaps, particularly in data generation and use, which were validated by the Jacobs Foundation's Evidence Navigation Journey (ENJOY) framework. To address these gaps, the Fellows prioritized the following questions to explore during the LEAP Sprint.

1. Based on the available ASER data, what insights can be gained regarding the effectiveness of the ENAd program? How can these insights inform the content and implementation of the program?

2. What are the limitations of the current data and the open questions that remain? How can data collection be refined to address these gaps?

3. How can data collection and utilization processes be optimized and systematized to enhance the efficiency and effectiveness of data-driven decision making?

4. What specific guidelines and templates can be developed to enhance the consistency and usability of data collection, analysis, and reporting?

Solution summary & next steps

Deliverable 1:

This deliverable provides FAAQ with concrete advice on how to grow their ambition to become the leading Educational Solutions Lab' In Colombia. To be an effective education solutions and evidence lab, it is essential to use data and insights for continuous improvement and informed decision-making. This involves actively generating evidence that helps FAAQ and others adapt programs to the unique needs of Colombia. Deliverable 1 provides guidance on analyzing and visualizing current ASER data, as well as on enhancing future data acquisition to generate actionable evidence. Specifically, Deliverable 1 provides detailed guidance on different ways to analyze and visualize data that FAAQ has previously collected during their implementation of their ENAd program. The analyses, findings, and recommendations that are presented are meant as an example that can be transferred to other data sets that FAAQ may be collecting in the future and they grow as an Educational Solutions Lab'.

Deliverable 2:

Deliverable 2 provides FAAQ with a practical toolkit to strengthen data use across the program cycle. It consists of two integrated tools: the Evidence-to-Action (E2A) Framework for planning MEL strategies, and the E2A Tracker for documenting data collection, analysis, learning, and follow-through on reporting and taking action. Together, the tools give life to the **recommendations in Deliverable 1** by making it

easier for staff to collect the right data, ask the right questions, and follow through on what the evidence reveals.The toolkit reflects FAAQ's emerging identity as an Evidence Lab by supporting systematic, intentional, and adaptive learning at the program level.

Deliverable 1

Retrospective analysis & guidelines for future data acquisition: Generating actionable evidence

The goal of this report is to advise FAAQ on steps that will lead them to becoming the leading 'Evidence Lab' In Colombia. To be an effective education solutions and evidence lab, it is essential to use data and insights for continuous improvement and informed decision-making. This involves actively generating evidence that helps FAAQ adapt programs to the unique needs of Colombia. In this deliverable, we provide guidance on analyzing and visualizing current ASER data, as well as on enhancing future data acquisition to generate actionable evidence. The below guide to analyzing ASER data is meant as an example that can be transferred to other data sets that FAAQ may be collecting in the future as they grow as an 'Evidence Lab'.

1. Analysing current ASER data

The first chapter aims to explore the effectiveness of the ENAd program by analyzing available ASER data, guided by the following questions:

- *Based on the available ASER data, what insights can be gained regarding the effectiveness of the ENAd program?*
- *How can these insights inform the content and implementation of the program?*
- *What are the limitations of the current data and the open questions that remain?*

In section 2.1, we describe a few considerations regarding the data analyses, while in section 2.2 we outline the steps to perform a number of key analyses. In 2.3, we end with conclusions and open questions.

1.1 Analysis considerations

To determine the best way to analyze the current ASER data, it is important to consider the questions that you would like to address, the level of thoroughness you are aiming for, the type of data that you are working with, and the software that you would like to use. Each of these points is described in more detail below.

1. **The questions you aim to address *and* that can be answered using the available data.** The current ASER data may provide valuable (though limited) insights into students' progression in both literacy and numeracy throughout the ENAd program, as well as the factors influencing this development, such as the duration of implementation. Yet, it is important to keep in mind that, without a control group, we cannot investigate whether students progress faster than in regular, business as usual, schooling. This makes it difficult to draw strong conclusions about the program's effectiveness. Furthermore, it is important that while the brevity of the ASER is a real pragmatic advantage, it is also a limitation from a measurement perspective, as data on student literacy and numeracy knowledge and skills is limited to only a few items per measurement time point.

2. **The level of thoroughness and detail you are aiming for.** If you want to make claims about changes in students' literacy or numeracy over the course of the ENAd program, or the factors influencing it, advanced statistical methods might be necessary. However, if you want to get some idea about general trends for internal use, descriptive statistics and data visualisation may be sufficient.

   A note on statistical significance

   It is important to realize that your sample size is very large (> 1000 students), which means that you have a lot of statistical power. As a result, the effects assessed by statistical tests will often be statistically significant, even if the effects (e.g. differences in ASER scores between time points) are too small to be meaningful.

   When studying differences between schools, the opposite problem may arise. Because there are only a few schools in the current samples (< 20), findings may not be significant, even if a real effect exists.

   In both cases, it is valuable to reflect on the meaningfulness of the improvement.

   How to determine if the improvement is meaningful?

- Report the effect size (also see this tool for visualizing effect sizes), which can be selected as an output when running a given statistical analysis. For example, if you run a t-test you can obtain the effect size as a measurement of the difference in ASER performance between, for example, two measurement time points. The effect size is an indication of the strength of a finding, which unlike a difference in means takes into account the variability in your data, and can help better understand the magnitude of any effect that you are measuring. For example, you may want to better understand the effectiveness of ENAd in two different samples from two different locations in Colombia. You run a statistical test to assess whether the improvement of students across baseline, midline and endline was statistically significant. However, when you calculate the effect size you find that the size of improvement in one sample was 0.1 of a standard deviation and in the other sample you find that the size of improvement (as measured by the effect size) was 0.5 of a standard deviation. While in both groups the effect of ENAd was significant the effect size tells you that the magnitude of the effect of ENAd was much larger in one sample compared to the other. In this way effect size can give you very important information to better understand the magnitude of any effect you want to measure over and above its statistical significance.
- Visualize your findings and determine whether you have reached a predetermined 'minimum threshold', for example the percentage of lagged vs. non-lagged students.

Taking into account 'nested data'

Ideally, analyses should account for the fact that students are 'nested' within schools. This means that students within the same school are not entirely independent as they share common characteristics. To properly account for this structure, multilevel analysis is often used to separate the variation at both the student and school levels (in larger samples you can also add classroom as another level of your multilevel model). This approach requires advanced skills in statistical analyses.

3. **The type of data that you are working with.** ASER data is ordinal**.** This means that skill levels are ordered from more basic skills to more complex skills, but we don't know exactly how much 'better' one level is compared to the next. Because of this, calculating averages or using parametric statistical tests can be misleading because they assume equal difference between values. Instead, non-parametric tests are better suited for analyzing ordinal data since they

focus on rankings rather than exact values. Examples of such tests and appropriate descriptive statistics are provided in section 1.2.1.

Aggregating student data per school

When studying the effects of program, implementation, or school-related factors, you might consider summarizing student data at the school level, as most of these factors operate at that level. There are several ways to aggregate data at the school level, so it's important to choose the outcome variable that best aligns with your objectives (e.g., the percentage of lagged students or the percentage of students who reach a certain level). When you summarize the data in this way, they are no longer ordinal, allowing the use of parametric tests. Examples of such tests are provided in section 1.2.2.

One important consideration when performing analyses at the school level is that the number of students per school can vary significantly. This has two potential challenges:

- In schools with fewer students, performance of an individual student has a greater impact on the overall score, making outliers more influential.
- Students in schools with fewer students have a disproportionately larger impact on the outcomes of your analysis, compared to students in larger schools.

A weighting factor could help balance these problems, but it might also add complexity to the analysis. Instead, an alternative approach is to exclude schools with very few students. There is no strict rule for the minimum number of students per school. A minimum of 15-20 students might seem reasonable, but the choice should also consider the average school size in the dataset to maintain representativeness.

4. **The software that you would like to use.** Different tools (e.g., Excel, JASP, JAMOVI, SPSS, or R) differ in the complexity of the analyses that they offer, as well as their ease of use. Furthermore, whereas some tools are free (such as R or JASP), others come with certain costs, so the choice depends on your analysis needs, skill level, and budget.

   We recommend using a combination of excel and JASP. JASP (https://jasp-stats.org/) is open-source software for statistical analysis, which has a user-friendly interface and allows a variety of statistical methods.

1.2 Analysis steps

1.2.1 Main analyses: student achievement at baseline, midline, endline

Goal: Get insight into the levels that students reach at baseline, midline, and endline. What proportion of students are at each level? Do students seem to get 'stuck' at some levels? What is the level that most students are at?

Approach:

A lot can be learned by looking at descriptive statistics and plotting the baseline, midline and endline data in various ways. We recommend looking at:

- The median (i.e., the middle value when the data are ordered).
- The range (i.e., the range between the highest and lowest level).
- The interquartile range (i.e., the spread of the middle 50% of data).
- The frequency distribution (i.e., the percentage of students in each level).
- The percentage of lagged students (the outcome that is primary interest to you)

Although most of these measures can be found in excel, we focus on JASP here because of its intuitive interface and nicer visualisations.

Step 1. Create new variables that show (a) the highest level reached at baseline, midline, endline, and (b) if students were lagged vs. not lagged at baseline, midline, endline

In excel, add 3 columns to show highest level reached at baseline, midline, endline (LB, LM, LF)
Excel formula =@IFS(L5=1;0; M5=1;1; N5=1;2; O5=1;3; P5=1;4; Q5=1;5)
(L5-Q5 are the columns representing the levels)

To indicate whether a student is lagged vs. not lagged at baseline, midline, endline, add 3 more columns
Excel formula =IF(AM5=5;1;IF(AM5<5;0;""))
(AM5 is the column with highest level for each student, which was created in the previous step)

Copy data to JASP by selecting the table including the headers and pasting it into JASP. Do not include the summary row at the bottom.

Step 2. Descriptive statistics and basic plots.

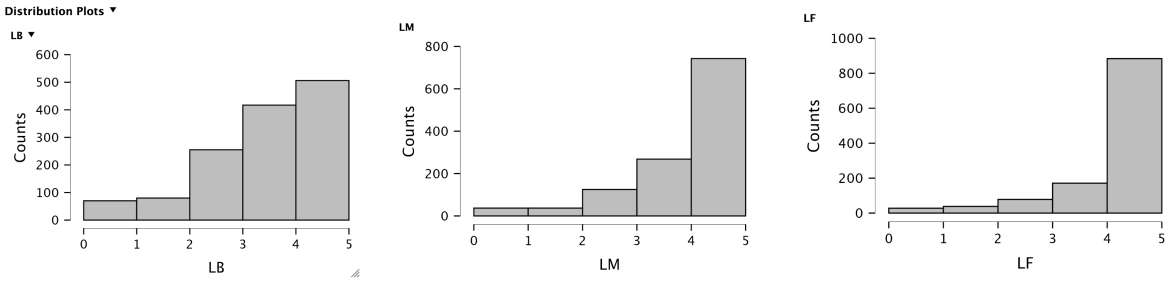
2a. Distribution of levels across all students

One of the first things you may want to visualize is how all students performed at the three measurement time points (baseline, midline, endline).

In JASP select *Descriptives*. Once the Descriptive window opens you will see on your left side a box that lists the variables specified in your data set.
You will then be able to transfer those variables you want to obtain the descriptives for to the empty box. Let's take the example of the 'Risaralda 2004 1 literacy' data set. You can then select variables 'LB' 'LM' and 'LF' and transfer them into the box on the right side.

This will immediately provide you with a table of descriptives in the right output window. If you expand the statistics button immediately below the left window you can plot different statistics such as mode or median by selecting these statistics to be included in your table. From this table you can see, for example, that the median level achieved by students increases across the three time points.

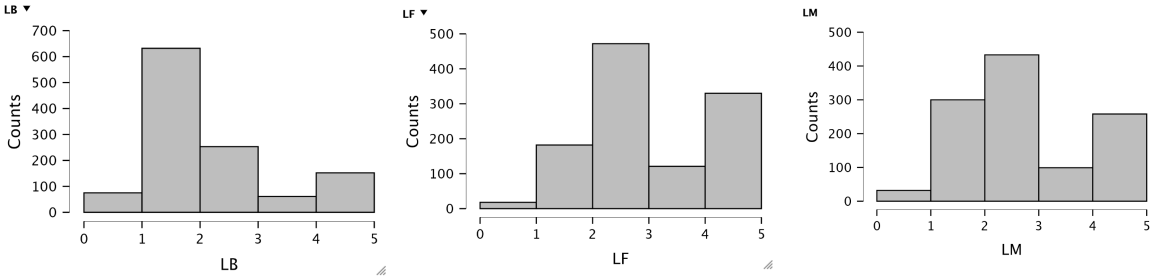
However, this table is perhaps not the best way to describe your data. Instead you may want to visualize the distribution of students at each level. For this you need to expand the 'basic plots' menu on the left hand side. Then select 'distribution plots'. This will then output the distribution of students at each level on **ASER Reading** separately for baseline (LB), midline (LM) and endline (LF):



From these distribution plots you can immediately see that at all three time points the ASER Reading data is left skewed (that is the weight of the data is distributed towards the higher ASER reading levels). In addition you can see that across the three time points more and more students reach higher levels. This is a useful way of

visualizing the change over time and seeing that more and more students reach higher levels.

Below is distribution of students at each level on the **ASER Numeracy Data** separately for baseline (LB), midline (LM) and endline (LF):



It is immediately apparent that the distributions of the Numeracy Data look different from the distributions of the literacy data at each time point. It is clear that fewer students end up at the highest levels and that many students appear to be struggling with the transition from level 3 (subtraction) to level 4 (division). Once they master level 4, they move relatively quickly to level 5 (word problems). Please see section 2.2 below on 'Assessment considerations' for further discussion of this finding and possible future steps.

2b. Distribution of levels, split per grade for ASER literacy (left panel) and ASER numeracy (right panel)

Descriptive Statistics

	LB	LM	LF
Valid	1328	1210	1199
Missing	82	200	211
Mode	5.000[a]	5.000[a]	5.000[a]
Median	4.000	5.000	5.000
IQR	2.000	1.000	1.000
Range	5.000	5.000	5.000
Minimum	0.000	0.000	0.000
Maximum	5.000	5.000	5.000
25th percentile	3.000	4.000	4.000
50th percentile	4.000	5.000	5.000
75th percentile	5.000	5.000	5.000

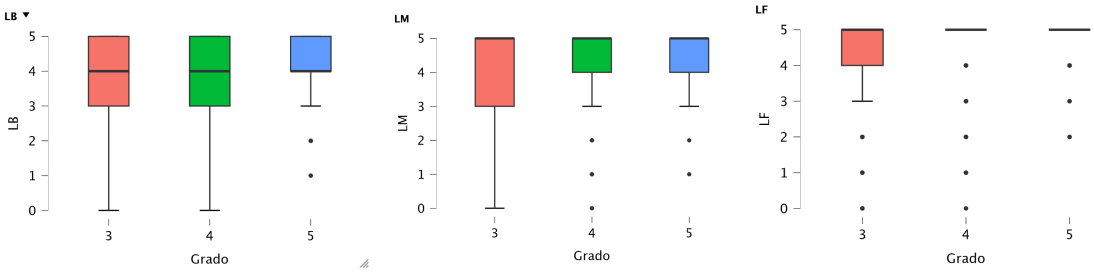
Descriptive Statistics

	LB	LM	LF
Valid	1098	1060	1050
Missing	84	122	132
Median	2.000	3.000	3.000
IQR	1.000	2.000	2.000
Range	5.000	5.000	5.000
Minimum	0.000	0.000	0.000
Maximum	5.000	5.000	5.000
25th percentile	2.000	2.000	3.000
50th percentile	2.000	3.000	3.000
75th percentile	3.000	4.000	5.000

In addition to describing the data of an entire sample, you may want to understand how students are distributed across levels within each of the grades that have run your program in. In order to do so you would go back to Descriptives, enter your Baseline (LB), Midline (LM) and Endline (LF) variables just like you did above, but now also enter your grade variable (Grado) in the box on the bottom of the right hand side that says 'Split'. Please note that for the purpose of this example, we have selected grade 3,4 and 5 only.

As in 2a above you will immediately see a table with all the descriptives. However, this table now contains a lot of data as it is split by the grade of the students and thus is hard to digest.
Therefore, we would recommend that you go straight to a visual representation of your data by grade.

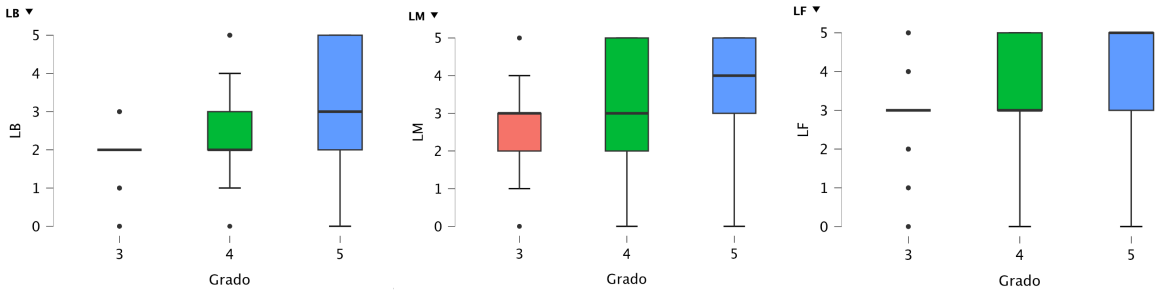
To do so, expand the 'Customizable plots' menu on the left hand side. In order to better understand the distribution of data within each grade and time point we would recommend selecting 'Boxplots''. Also check the 'Use color palette' to enable you to better visually distinguish between different grades (you can see that you can select different color pallets - for the purpose of this example we selected the pallet 'ggplot2' from the popular ggplot toolbox in R - you can play around with the different pallets to see which one best serves your purposes). Once you have selected 'Boxplots' and 'Use colour palette' you will see three graphs generated on the right hand side that show you the distribution of students in each grade for **ASER Literacy** separated by the testing time point:



In a boxplot, the box represents the 'interquartile' range and indicates where 50% of your data are distributed. The thick line represents the median and the 'whiskers' represent the top and bottom 25% of your data. The dots represent outliers which you can label by selecting 'label outliers'. This is useful as you can then inspect these outliers and perhaps check against notes as to anything unusual about these students.

As you can see from the three boxplot panels above, students in each grade improved in their performance, as you can see that the thick black line (median performance within each grade) moves 'up'. Furthermore you can see that the variability in the data decreases as a function of assessment time point. Indeed at the endline you can see that the median for students in grades 4 and 5 is 5, which is the ceiling on the **ASER Literacy** test. In grade 3 there is still some variability, as might be expected given that these students are the youngest learners.

In the following you will see the boxplots showing you the distribution of students in each grade for **ASER Numeracy** separated by the testing time point:



As you can see from the above (and consistent with the histograms showing the distribution of data across grades above), you can see that the ASER numeracy data is more variable across measurement times (LB, LM, LF) and grade (3,4,5). Consider the midline data, here you can see that in grade 4 most students are still at level 3 on the ASER numeracy test and the variability is substantial as can be seen from the box and whiskers. Furthermore at endline (LF) the variability remains substantial especially compared to the ASER literacy data plotted above.

Boxplots are excellent for your internal use to better understand the data. They are probably not the best ways of showing your data to your board or educators as they contain a lot of statistical information.

Step 3. Significance testing using nonparametric tests.

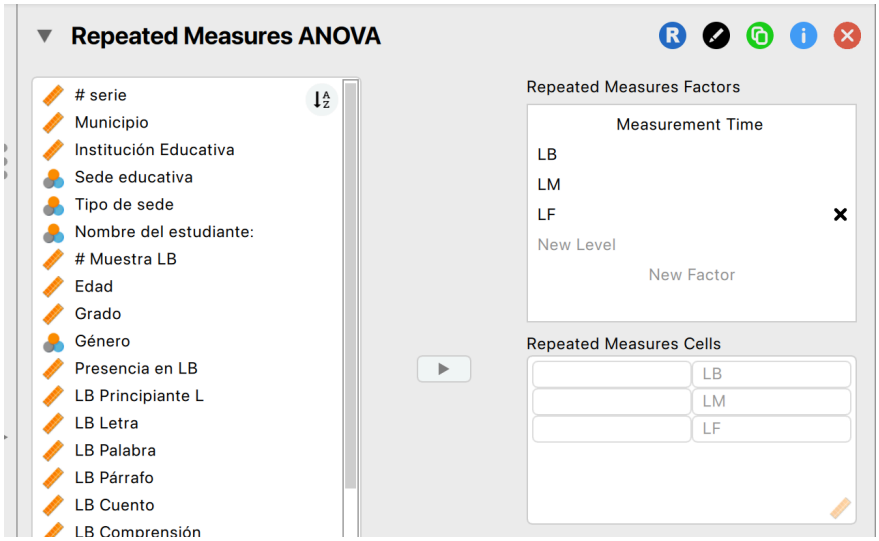
Beyond descriptive data you may want to know whether the students improved significantly across your three measurement time points. As stated the ASER data are non-parametric and therefore we need to run non-parametric statistical analyses. In order to do so do the following in JASP:

Go to ANOVA on the top bar - select ‘Repeated Measures ANOVA’ - we are selecting this as we are looking within students across time and thus we are analyzing repeated measures data.

This will open the Repeated Measures ANOVA Window. The first thing you have to do here is to label your factor and its levels. In your case your factor is Measurement Time with three levels: Baseline, Midline and Endline.

First click on ‘RM Factor 1’ in right sided window and change the label to ‘Measurement Time’ .

The Change Level 1 below to ‘LB ’ for Baseline, Level 2 to ‘LM’ for Midline and Change ‘New Level; to ‘LF’ for Endline. Now you have specified your statistical model and it should look like this:



Now drag your variables for your three (LM, ML and LF) measurement time point where it says ‘Repeated Measures Cells’

This will immediately output a parametric analysis called ‘Repeated Measures ANOVA’. However, given that we are dealing with non-parametric data we are going to ignore this and instead go to the very last expandable menu on the left hand side that is labelled ‘Nonparametrics’.

Once open, drag your Factor ‘Measurement Time’ over into the window that is labelled RM Factor.

This will then output a so-called ‘Friedman Test’ on the right hand side. This is a repeated measures test for non-parametric data. The output includes the p-value (significance) and the Kendall’s W, which is a measure of the effect size. Kendall's W

can vary between 0 and 1, where values closer to 1 indicate that the effects are more consistent across students.

Nonparametrics

Friedman Test

Factor	X^2_F	df	p	Kendall's W
Measurement Time	596.669	2	< .001	0.290

As you can see in the table above, the Friedman test shows that there is a significant effect of measurement time. That means that the different time points are significantly different from one another. However, it just tells us that there is an effect of time on levels, not whether the 3 different levels are each statistically different from one another. If we want to know this we have to select 'Conover's post hoc test' from the left hand side. This will then output the following:

Conover Test

Conover's Post Hoc Comparisons – Measurement Time

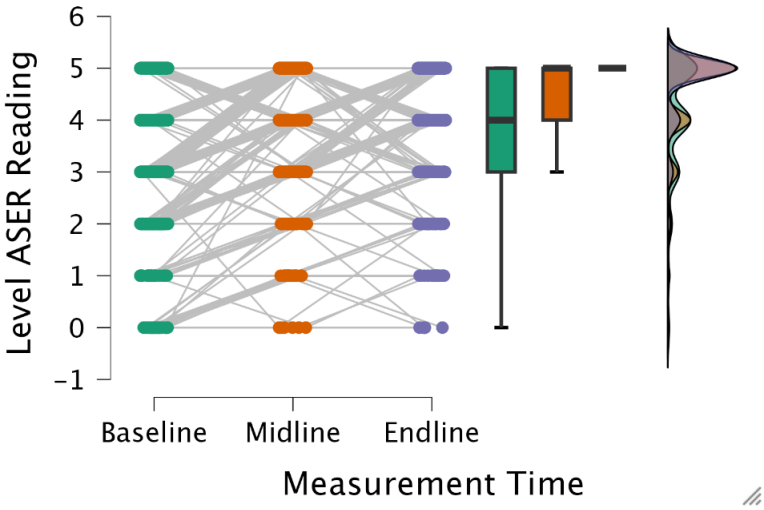
		T-Stat	df	W_i	W_j	r_{rb}	p	p_{bonf}	p_{holm}
LB	LM	19.853	2056	1646.500	2156.500	−0.778	< .001	< .001	< .001
	LF	28.202	2056	1646.500	2371.000	−0.878	< .001	< .001	< .001
LM	LF	8.350	2056	2156.500	2371.000	−0.602	< .001	< .001	< .001

Note. Grouped by subject.
Note. Rank-biserial correlation based on individual signed-rank tests.

This post-hoc test allows you to see the differences between each of the three levels. So for example you can look at the first line which compares LB to LM, the second line is the comparison between LB and LF and the last line represents the comparison between LM and LF. If you look over towards the right hand side of this statistical output you can see that each of the comparisons are highly significant (even when controlling for multiple comparisons, which are represented by Pbonf and Pholm). So this tells us that performance on ASER was significantly different between a.) LB and LM, b.) LB and LF and c.) LM and LF.

We can then go on to visualize this data. Here we recommend using the Raincloud Plots as they provide a very comprehensive visualization of your data.

To do so, on the left hand side, scroll down and expand the menu 'Raincloud Plots', drag your Factor 'Measurement Time' to 'Horizontal Axis' and provide a label for the y-axis on the left hand side (Label y-axis). This will generate your Raincloud plot:



This provides you with a very comprehensive visualization of your data. *Please note that you can run these same analyses with the ASER Numeracy data.* You can see the individual trajectories of students on the left hand side. The boxplots for each time-point in the middle (see description of boxplots above) and the distribution of the data at all three points across levels on the far right hand side.

1.2.2 Aggregating data per school

Goal: Get insight into differences between schools in the proportion of students who reach a certain level, and study the effect of differences in program implementation (e.g., number of hours).

Approach: First, aggregate data at the school level in excel, using pivot tables. Next, analyze the data in JASP using parametric tests. The example below is from the Risaralda 2024-1 cohort.

Please see Appendix 2.1 for how to rearrange the data so that it can be used in JASP

Step 1. Pivot table in excel with # kids at each level per school, including # hours and average grade (PLEASE SEE APPENDIX 2.1 for details)

Step 2. Copy pivot table & calculate percentages per school (PLEASE SEE APPENDIX 2.1 for details)

Step 3. Analyses in JASP

Important Note: The sample size, i.e., the number of schools in the 2024-1 Risaralda sample, is too small to conduct statistical tests (6 urban schools plus the average of all rural schools). However, we outline the analysis process so that it can be applied when more schools participate in the future. We recommend conducting these analyses when the sample includes approximately 30 schools or more, with each school having enough students to give you a reliable percentage.

Below, we present a way to study changes in the percentage of students who reach at least level 3. The same logic can be applied to study changes in the percentage of students who are lagged vs. not lagged or other outcome measures. We also describe how to study the effect of program, implementation, or school-related factors.

- Go to ANOVA > Classical > Repeated Measures ANOVA
- Add a level to the *Repeated Measures Factors* box in the middle and rename each level to Baseline, Midline, Endline. Rename 'RM Factor 1' to something meaningful, like 'Level3andUp'
- Add the variables with percentages to the *Repeated Measures Cells* box: B345pct, M345pct, F345pct
- School-level factors can be added to *Between Subject Factors* (if it concerns categorical factors such as intervention type A vs B) or *Covariates* (if it concerns continuous factors such as the amount of instruction hours).
- Scroll down to *Display*, and check the boxes *Estimates of effect size* and then *omega* ω^2 or *partial omega* ω^2. This measure shows how much achievement changes across time, by quantifying the proportion of variance in achievement that is explained by assessment time (Baseline, Midline, Endline). Partial ω^2 controls for other factors, if applicable.
- On the right appears a table with statistics. The first line in the upper table (Level3andUp) shows to what extent achievement levels changed from Baseline to Midline to Endline. The **p-value** shows you whether the difference between timepoints is significant (if $p < .05$). The partial omega squared shows the **effect size**. Partial omega squared is typically interpreted using the following guidelines: a small effect is around 0.01 (explaining approximately 1% of the variance), a medium effect is around 0.06 (explaining about 6% of the variance), and a large effect is around 0.14 or higher (explaining 14% or more of the variance).
- If you have added school-level factors (for example the amount of instruction hours), the second line in the table (Level3andUp * Average of Número de horas TOTALES) shows you to what extent changes in achievement are influenced by the factor of interest. Because of the low number of schools, we have not performed this analysis for the 2024-1 Risaralda sample.

- Scroll down to *Raincloud Plots*. Move the Factor ‘Level3andUp’ to *Horizontal Axis*. Three graphs appear on the right. The first graph shows the percentage of students that have reached at least level 3 for each school at each timepoint. The box plot in the middle provides summary statistics (median, quartiles, potential outliers). The graph on the right shows the overall distribution for each time point. You can change the range of values at the y-axis by clicking on the triangle next to ‘Dependent’, then ‘edit image’ and then go to y-axis.

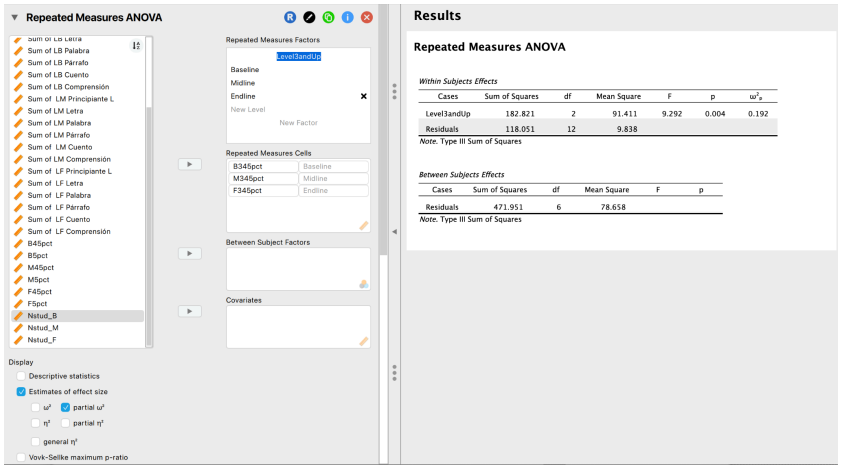
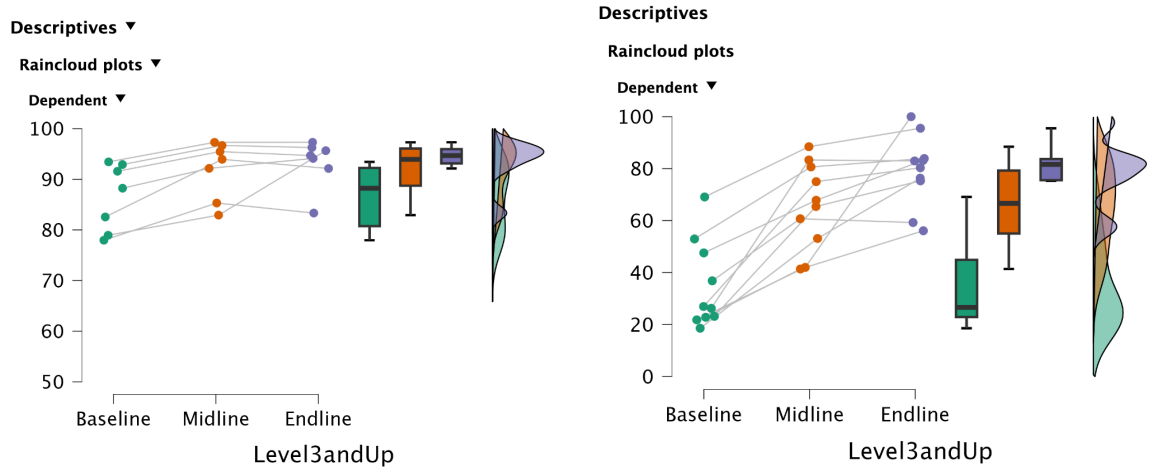
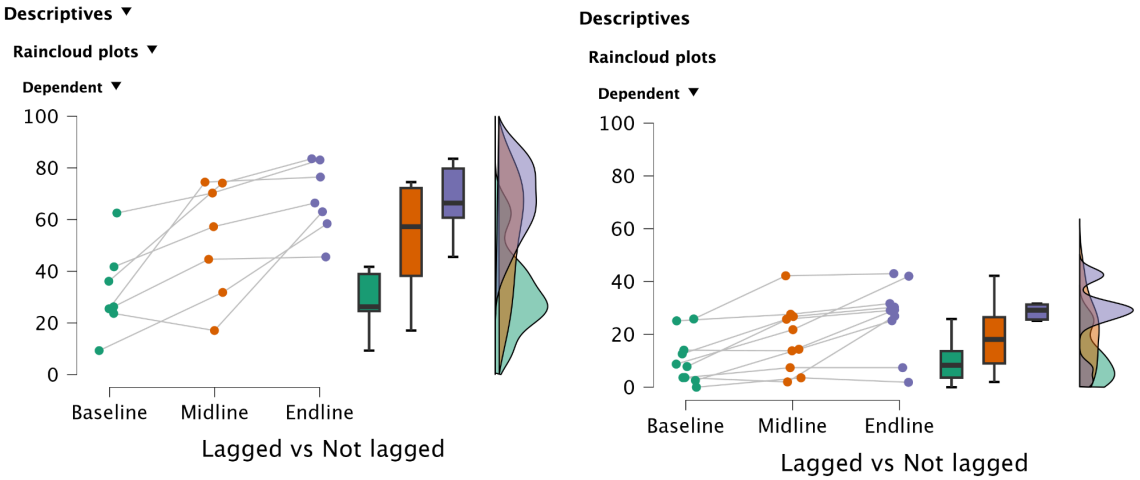


Figure X Repeated measures ANOVA in JASP. The findings on the right are from the literacy data. Note that these findings should be interpreted with caution due to the small sample size.



JASP output of school aggregated data for literacy (left) and numeracy (right) showing the percentage of students who have reached at least level 3 at Baseline, Midline, and Endline.



JASP output of school aggregated data for literacy (left) and numeracy (right) showing the percentage of students who are not lagged in literacy at Baseline, Midline, and Endline.

1.3 Insights from ASER data

Insights gained

What we can see from the example analyses above, is that both at the individual level and at the school level there is evidence for significant improvement in student ASER scores across the three measurement time points. What we can also see is that there is considerable variability between students and between schools. This variability is interesting and could be further explained by including other variables linked to the fidelity of implementation, student and school-level socio-economic status and teacher variables (engagement, preparation to teach ENaD etc.).

Importantly, we can see quite striking differences in students' performance on ASER Reading and Math between assessment time points (baseline, midline, endline). While we can clearly see progress in reading, the same is not true for numeracy where students seem to get stuck on the subtraction item and do not progress in the same way.

We looked at this in a little more detail. when we compute new variable in JASP (using the 'computed with drag and drop' function in JASP- see Appendix 2.2 for details) that enumerates students that are at the subtraction level at both midline and endline we find that 279 students are at the subtraction level at both midline and

endline. Furthermore, 116 students are at the level of double digit naming at both midline and endline. This suggests that students are struggling to move between the different ASER Numeracy Levels. Compare this to the literacy data. When we compute which students are at the ASER Word ('Palabra') level at both midline and endline we find there are only 17 students who are at this level at both times. When we consider the students that are at the ASER paragraph ('Parrafo') level at both midline and endline we find that there are 38 students who are at this level at both time points. Therefore, there is clear evidence from our analysis that more students get stuck at the lower ASER levels for numeracy compared to literacy.

We also see a lot more variability in student performance on numeracy compared to literacy. A noteworthy finding in the area of literacy is that many students achieve the highest level at baseline, particularly in grade 5. This raises the question of whether a school-wide implementation of the literacy program is necessary, or if it would be more effective to target only those students who require additional support. It may also be worth exploring whether the program could be expanded to include more challenging texts, providing advanced students with further opportunities to develop their reading comprehension skills.

Limitations and future directions

One of the limitations is that there is very little information about the students and schools other than the ASER data. This means that it is hard to account for the variability observed. In future FAAQ might consider collecting more demographic data such as SES at both the student and school level as well as variables related to the implementation and the fidelity thereof. Also, relying on a single measure to characterize student achievement provides a rather limited picture (see Assessment Considerations in 2.2 below). With respect to the school level data, these should be interpreted with caution since the number of students per school varies substantially and thus some school level data may be unduly influenced by individual students.

2. Becoming an Evidence Lab

With the goal of becoming an Evidence Lab in mind, it is crucial to take a step back and consider the type of evidence you are collecting and its intended purpose. This chapter helps you with that. In section 2.1, we describe how data can be used in different ways, depending on whether the data are used to adjust instruction for individual students (student-level goal), improve guidance for schools during

implementation (school-level goal), or refine the program as a whole (program-level goal). Taking this distinction in mind, section 2.2 provides a number of recommendations to improve current numeracy and literacy assessments. Section 2.3 focuses on program-level goals, highlighting a number of key questions that may help improve the ENAd program.

2.1 Goals of data collection

Data can play a key role in improving the program—both **during implementation** (by monitoring student progress and making adjustments as needed) and **between cycles** (by analyzing what factors contribute to program success).

Data collection can serve different goals:

- **Student-level goals.**
  - When: During an implementation cycle.
  - Goal: provide the right level of instruction for each child.
- **School-level goals.**
  - When: During an implementation cycle, or between cycles if the same school participates again.
  - Goal: offer the appropriate level of guidance to each school to help them better support their students.
- **Program-level goals.**
  - When: After a program cycle.
  - Goals: Improve the program's content and implementation to maximize its effectiveness. Assess the program's effectiveness.

It's important to distinguish between the different goals of data collection, as different types of data and analyses may be needed depending on the goal.

Example: If the goal is to group students according to their achievement levels (student-level goal), detailed data on specific sub-skills that they master may not be necessary. However, if the goal is to understand why some students get 'stuck' at a certain level (program-level goal), more detailed information becomes crucial.

2.2 Assessment considerations

Assessments are a critical component of any successful, evidence-informed, implementation of educational programs and interventions. The use of assessments,

such as screeners, curriculum-based measures and standardized tests, can serve multiple purposes in the implementation and evaluation of novel educational programs and interventions. As described in section 2.1, assessments can be used to evaluate the efficacy of a new educational program to decide whether to continue and scale its implementation. Yet, they can also be used to monitor student progress and to allow educators to adjust the program in response to the outcome of the assessment. This can take the form of adjusting instruction for individual students or grouping students according to their performance on assessments.

Current Assessments used by FAAQ:

Presently, FAAQ is using the ASER Numeracy and Literacy assessment tools. These measures each have 5 items that are organized in increasing difficulty. Students are assessed three times: baseline, midline, and endline. Students are assessed for the level of their performance on the ASER Literacy and Numeracy assessment tools. Each of the items represents a level and thus there are 5 possible levels. In alignment with TARL, FAAQ uses the assessment to group students who are at a particular level to target the instruction at their level.

FAAQ does not use other student-level assessment tools.

Advantages*:*

- The ASER test has been used in multiple other TARL implementations and thus is a well-used and proven tool.
- The ASER test is extremely rapid and thus does not take away from valuable and limited instructional time.

Disadvantages*:*

- The ASER test only contains 5 items per test (literacy and numeracy) and thus only gives a very limited impression of the level that learners are at and does not assess all the skills they are being taught.

- The ASER test relies on the assumption that the jump from say Level 1 to Level 2 is the same as from Level 3 to Level 4 when it comes to the learning challenge of moving between levels. This may not be the case. For example, the jump from naming single and double digits to subtraction is arguably a lot bigger than the jump from division to word problems. To be more specific, a child that can name single and double digits may not have a good understanding of what these symbols mean, how they refer to quantities and how they can be used to carry out operations (arithmetic).

Possible additions/modifications:

To gain a better understanding of why some students remain stuck at a particular level, it may be useful to consider some of the following recommendations:

Recommendations specific to Numeracy:

- Consider adding in additional ASER questions. For example, add in the addition item for the numeracy ASER test, as this may help to see students who are moving from double digit naming to addition, but do not yet understand subtraction. Furthermore, adding the multiplication test will help determine whether students are truly stuck at the subtraction level, or if they are beginning to learn multiplication before having mastered division.

- Consider assessing not only number naming but also number comparison. Number naming is a procedural skill that does not necessitate students' understanding of the numbers they are reading. Number comparison taps into their ability to retrieve the quantities associated with the symbols and thus taps into their emerging conceptual understanding of numbers.

Recommendations specific to Literacy:

- Upon reviewing the literacy data, we observed that many students who answered the comprehension questions correctly at baseline did not do so at midline or endline (data not shown in the current report). While the FAAQ team describes this decline as "learning loss," it is likely that the drop in performance can be (partly) explained by natural fluctuations in student engagement and performance over time. However, this pattern also raises questions about the reliability and validity of the assessment. We recommend a thorough review of the reading comprehension questions to ensure they target key aspects of the text—rather than minor or easily forgotten details—and require a true understanding of the texts to answer correctly.

Recommendations for both Numeracy & Literacy:

- Consider implementing the Early Grade Reading and/or Early Grade Mathematics Assessments (EGRA & EGMA) in one or more of the following ways:

- Give these tests to students who are not moving up in terms of their learning levels by midline to help you better identify which aspects of literacy and numeracy

they are struggling with and then target instruction to the weak literacy and numeracy skills revealed by the EGRA/EGMA.

- Assess students using EGRA and EGMA at baseline and endline to get a better idea of what skills students are learning and what they might still be missing. In addition, this approach would provide a much more detailed assessment of the efficacy of the program, ideally in comparison with a control group.

- It should be noted that this would require additional resources and time as the EGMA/EGRA tests take longer to administer and need to be administered one-on-one.

- Since the current assessment is entirely oral, it may be worth considering a (partly) written format that can be administered to the whole class simultaneously. Written tasks place fewer demands on working memory and offer greater efficiency, saving teachers time. They also allow for the inclusion of more problems at each level—enhancing reliability—and make it feasible to assess a broader range of levels, providing deeper insight into students' abilities.

2.3 Studying the program's effectiveness and areas for improvement

In this section, we outline strategies to assess program effectiveness and identify areas for improvement (program-level goals). Before collecting and analyzing data, it is important to consider what information will provide the most valuable insights into your program's effectiveness and areas for improvement. Here, we outline a number of key questions, and describe what data would be needed to answer these questions and how results would inform decision-making. We focus on quantitative data about student achievement. However, to enhance the evaluation of the program's content and implementation, these data should be triangulated with other data sources that FAAQ already collects, such as focus group discussions, surveys, and classroom observations, as well as new data sources like interviews.

1. **How do students' foundational literacy and numeracy skills develop over the course of the ENAd program?**
    - What is the proportion of students at each level when they begin the program?
    - What is the rate at which students progress through each level?
    - What percentage of students are no longer lagged after the program?

    Required data: Student achievement levels at baseline, midline, and endline (e.g., ASER data or EGMA/EGRA assessments).

How findings may inform decision making: Insight into the levels students reach at baseline, midline, and endline can guide adjustments to teaching strategies and assessment methods to better align with students' needs. For example, if the findings show that many students are stuck at a particular level, this could suggest the need for improved teaching strategies and/or a more fine-grained assessment method with additional substeps.

How findings cannot be used: These data cannot be used to draw direct conclusions about the effectiveness of ENAd. Because there is no comparison group, the observed improvements may also be caused by factors not related to ENAd (such as schooling in general). To properly assess the effectiveness of the program, a randomized controlled trial (RCT) is recommended (see question 3).

2. **Which factors influence learning gain?** There are several factors that can influence learning gains, which can be grouped into four broad categories:
    - Program-related factors (e.g., content of the lessons, instructional approach, formative assessment)
    - Implementation-related factors (e.g., teacher training, mentor visits, duration of program)
    - Population-related factors (e.g., socioeconomic status (SES), baseline achievement)
    - School-related factors (e.g., school leadership and support, willingness of teachers, teacher collaboration)

    Required data: Student achievement levels at baseline, midline, and endline (e.g., ASER data or EGMA/EGRA assessments) and factors that might influence learning gain. Program- or implementation-related factors should ideally be studied experimentally using an RCT. For instance, you could select a few schools to test a new version of the program (e.g., a new instructional approach) and compare their outcomes with those of similar schools receiving the regular program. Then, you would evaluate whether the schools with the new program perform significantly better than those with the regular version.

    If an RCT is not feasible, valuable insights can still be gained by analyzing naturally occurring variations across schools. It's important to keep in mind that these variations should not overlap with other relevant factors (e.g., schools that receive more mentor visits are often urban schools), as this could complicate attributing effects to a single cause.

    How findings may inform decision making:

- A better understanding of implementation- and content-related factors can inform overall program improvement. This applies particularly to factors that are within your control, such as the length and content of teacher training, the frequency of mentor visits, or the type of support provided.
- A better understanding of school- and population-related factors that affect program success may signal the need for additional support for specific schools or student populations.
- A better understanding of school-related factors may also help to prioritize partnerships with schools where the program is most likely to succeed and develop strategies to support schools facing implementation challenges.

3. **Does the implementation of ENAd lead to significantly and meaningfully larger improvements in foundational learning outcomes than those achieved through typical education?**

   Required data: Student achievement levels before and after the program (e.g., using ASER or EGMA/EGRA assessments). Since ENAd is expected to lead to improvement, it is critical to examine whether the improvement is larger than when students receive typical schooling. This would ideally require an RCT.

   Many organizations choose to hire an independent organization to carry out the RCT, as their results may be regarded as more credible by funders and other stakeholders. Another potential benefit is that external evaluators also often bring valuable expertise in research design, data collection, and analysis. However, hiring an independent organization can be quite resource-intensive, so this step is typically taken once the program has been sufficiently refined and there is confidence in its effectiveness. In the meantime, conducting an internal RCT can be highly valuable, providing early insights into the program's effectiveness. More specifically, it can help refine the intervention and identify areas for improvement, laying the groundwork for a more rigorous, externally validated trial in the future.

   How findings may inform decision making: If there is substantial improvement, this may justify scaling up the intervention, securing further funding, and advocating for policy adoption. If improvements are not as large as expected, program adjustments may be needed.

Deliverable 2

Deliverable 2: Evidence to Action Framework

Our Approach

As we began mapping the program flow and data collection journey, it became increasingly clear that we need to start from the organization's why. Without first anchoring in the broader goal the organization is trying to achieve, any process or data decisions risk being reactive or misaligned.

So instead of jumping straight to process mapping, we stepped back to ask:

- *What is the broader goal FAAQ is working toward?*
- *How does this program fit within their Theory of Change?*
- *What are the intended outcomes, and how can we reverse-engineer the program flow and data collection strategy to support those?*

By rooting your work in your broader vision and long-term goals, we're better positioned to:

- Align activities with outcomes – ensuring each part of the program is contributing meaningfully to your mission.
- Collect the right data at the right time – asking questions that matter, at the moments they matter most.
- Translate insights into action – ensuring that learning loops are built in, so you can course-correct, scale what works, and show evidence of success.

Here's how we're approaching the work:

1. Clarify the outcomes your Theory of Change aims to achieve
2. Map the program activities that directly support those outcomes
3. Define learning and evidence needs at each stage of the journey
4. Create a process flow and program journey map that integrates these insights
5. Recommend a good-practice approach for ongoing data use and reflection

Align activities with outcomes:

Ensure your work is grounded in your why so that you are working towards a shared goal:

Start with the Theory of Change – Understand the long-term outcomes and core assumptions.
Map the strategic activities – What are we doing, and why?
Clarify intended outcomes at each stage – What should change for participants along the way?

Collect the right data at the right time

1. Develop an aligned process to ensure you are collecting, analysing and measuring the right data:
2. Define the data strategy – What questions should we ask, when, and what does good evidence look like?
3. Build the program journey map – Showing both participant flow and aligned data collection moments.

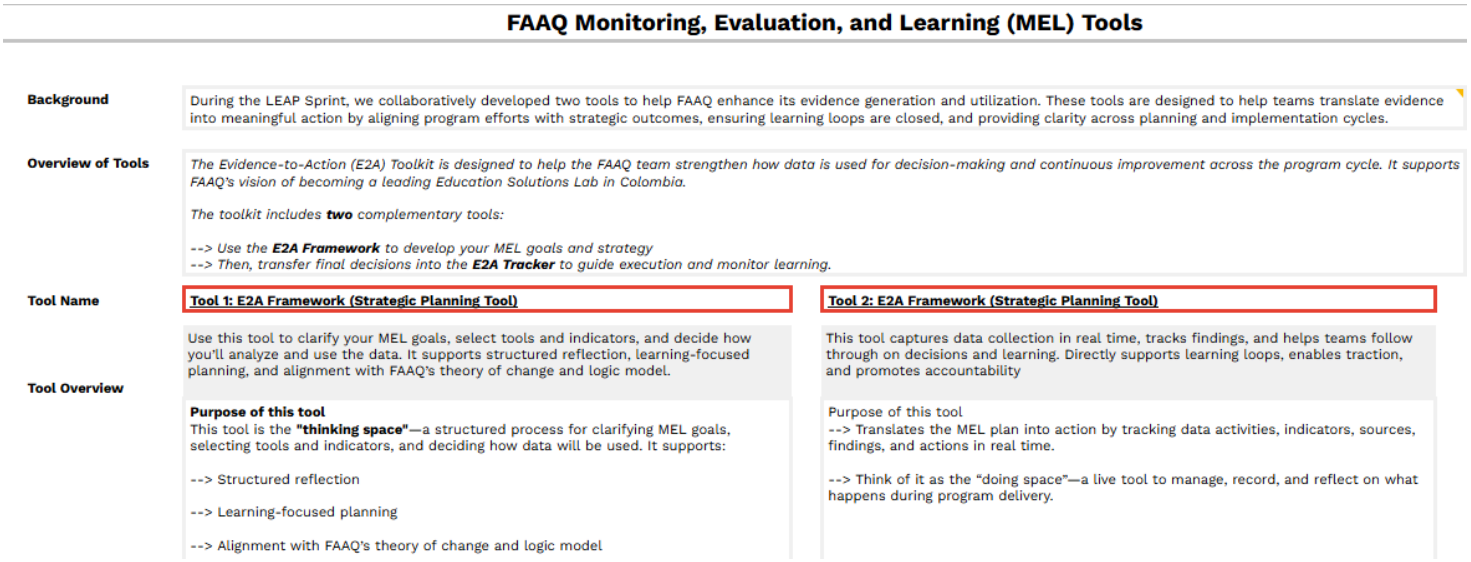
This approach allows you to:

- Design for impact, not just activity.
- Collect data that is useful and actionable.
- Ensure alignment between program design, execution, and evaluation.

This approach ensures that your data strategy is not just a checkbox, but a tool for continuous improvement, accountability, and meaningful impact.

3. Narrative Example: Applying the MEL Framework and Tracker with ASER

Please find the Deliverable 2 tool here:



FAAQ Monitoring, Evaluation, and Learning (MEL) Tools

Background	During the LEAP Sprint, we collaboratively developed two tools to help FAAQ enhance its evidence generation and utilization. These tools are designed to help teams translate evidence into meaningful action by aligning program efforts with strategic outcomes, ensuring learning loops are closed, and providing clarity across planning and implementation cycles.	
Overview of Tools	*The Evidence-to-Action (E2A) Toolkit is designed to help the FAAQ team strengthen how data is used for decision-making and continuous improvement across the program cycle. It supports FAAQ's vision of becoming a leading Education Solutions Lab in Colombia.* *The toolkit includes **two** complementary tools:* *--> Use the **E2A Framework** to develop your MEL goals and strategy* *--> Then, transfer final decisions into the **E2A Tracker** to guide execution and monitor learning.*	
Tool Name	**Tool 1: E2A Framework (Strategic Planning Tool)**	**Tool 2: E2A Framework (Strategic Planning Tool)**
Tool Overview	Use this tool to clarify your MEL goals, select tools and indicators, and decide how you'll analyze and use the data. It supports structured reflection, learning-focused planning, and alignment with FAAQ's theory of change and logic model.	This tool captures data collection in real time, tracks findings, and helps teams follow through on decisions and learning. Directly supports learning loops, enables traction, and promotes accountability
	Purpose of this tool This tool is the **"thinking space"**—a structured process for clarifying MEL goals, selecting tools and indicators, and deciding how data will be used. It supports: --> Structured reflection --> Learning-focused planning --> Alignment with FAAQ's theory of change and logic model	Purpose of this tool --> Translates the MEL plan into action by tracking data activities, indicators, sources, findings, and actions in real time. --> Think of it as the "doing space"—a live tool to manage, record, and reflect on what happens during program delivery.

[Tool 1 - Framework]

[Tool 2 - Tracker]

4. Recommendations

FAAQ is a young and ambitious organization with a passionate team that consistently punches above its weight. In less than two years, it has reached nearly 8,000 children and built partnerships across both public and private sectors. During this startup phase, what is most notable is that FAAQ prioritized building strong MEL capabilities and a culture of continuous learning from the outset. This commitment has allowed the team to rapidly adapt and refine its approach — a strength that will not only accelerate its impact but also help differentiate it within Colombia's education ecosystem.

As FAAQ gears up for its next phase of growth, the Fellows have put forward several recommendations for the organization to consider, accompanied by specific suggestions for the ENAd program.

4.1 Organization-level Recommendations:

- If FAAQ aims to become the go-to organization for adapting and scaling educational interventions in Colombia, it should consider ***establishing an organization-wide research and learning agenda aligned with this goal***. This agenda should prioritize the use of data from across its interventions to better understand the Colombian education system and identify effective strategies for driving change. This approach will not only enable FAAQ to generate the evidence needed to strengthen its programming but also solidify its position as a key thought leader and implementation partner in Colombia—one with deep, on-the-ground knowledge of the education sector and how to facilitate transformation within its unique context.

- Given FAAQ's capacity for rapid iteration and its growth plans, we see significant value in ***making its evidence-building and decision-making processes visible.*** Just as we ask students to show their work to clarify their thinking and provide support, FAAQ can document how evidence and feedback have influenced key choices related to organizational strategy and program design. By helping others understand where FAAQ has been and what it has learned along the way, the organization can build credibility, strengthen institutional memory, and foster alignment with internal and external stakeholders. This transparency will enable those outside the senior leadership

team to engage more effectively as thought partners and leverage past experiences instead of repeating mistakes.

- FAAQ can build on the E2A tools by ***turning the E2A framework and tracker into a simple system for capturing and sharing what the team is learning.*** Right now, the E2A tools help track progress and guide decisions. With a few small changes—like organizing insights by theme or saving them in a shared space—FAAQ can create a go-to place for lessons learned. This will make it easier for new team members to get up to speed, help everyone stay aligned, and make sure valuable knowledge isn't lost as the organization grows.

4.2 ENAd Program Recommendations:

- When selecting which assessments to include, it is important to ***distinguish between the different goals for which the data will be used***. A primary goal of the ASER assessments is to monitor student progress in order to provide appropriately leveled instruction for each child. Additionally, the data can offer valuable insights at the school level, such as identifying which schools may require additional support. Besides these insights that are particularly relevant during program implementation, data can also provide insights into the overall effectiveness of the program and areas of improvement from one implementation cycle to the next. Clearly defining these goals will help ensure that data collection is purposeful, targeted, and aligned with decision-making needs. This point is further elaborated upon in *Section 2.1* of the report.

- To better understand the factors contributing to program success, we recommend ***documenting differences between schools***—both in terms of contextual background and implementation practices—and examine the extent to which these factors explain differences in program effectiveness. This may yield valuable insights for further refining and strengthening the program. This point is further elaborated upon in *Section 2.3* of the report.

- We recommend ***maintaining a codebook*** to clearly define the meaning of each variable and document essential details related to each one. This practice will facilitate future analyses, ensuring that anyone reviewing the data can easily understand the context and methodology behind it.

- As FAAQ continues to strengthen its MEL approach for the ENAd program, it should consider how to ***meaningfully engage school and local stakeholders—ensuring they have ownership of the data, are accountable for***

results, and collaborate with FAAQ to improve student learning. To support this effort, FAAQ can explore the Data Wise Project at the Harvard Graduate School of Education, which helps schools, districts, and organizations build their capacity for collaborative data use and continuous improvement. Furthermore, efforts to build school and local capacity for evidence use and joint problem-solving align with the Jacobs Foundation's EdLab Alignment Framework, which emphasizes co-creation, strong relationships among evidence actors, and the need to build a culture of data use across all levels of the education system.

- FAAQ is aiming to ***strengthen its mentoring tool to generate more actionable evidence on instructional quality***. Two widely used, research-based classroom observation tools are the Danielson Framework for Teaching and the World Bank's TEACH Primary tool. Both have been implemented across multiple countries and school systems, and are positively correlated with improvements in instructional quality and student learning outcomes. These tools provide clear rubrics that define effective teaching, pinpoint strengths and areas for growth, and can be used for both professional development and teacher evaluations.

- In addition to exploring potential ways to modify and improve the mentoring tool itself, FAAQ may consider ***training teachers on how to use the classroom observation tool.*** Research shows that training teachers to use structured classroom observation tools creates a common language and shared understanding of what good teaching looks like, enabling them to better meet expectations and reflect—both individually and collectively—on their practice (Klette, 2023). In fact, districts that train teachers—not just evaluators—to use structured frameworks improve pedagogical knowledge, instructional alignment, and quality (Kane et al., 2011) and those that institutionalize peer observations and feedback are more likely to sustain instructional improvements over time (Ridge & Lavigne, 2020). In situations where building school-level capacity and fostering a supportive, collegial culture are challenging or out of scope for FAAQ, it may be worth considering how to embed the use of classroom video clips during trainings or mentorship meetings to help teachers gain similar benefits. Several studies indicate that having teachers analyze classroom video clips enhances professional learning, self-reflection, and instructional quality (Brouwer, Besselink, & Oosterheert, 2017) while also helping them become more analytical and student-focused (Sherin & van Es, 2009). Additionally, watching videos of teaching that do not feature themselves or their colleagues may encourage teachers to be more open to providing constructive criticism.

Appendix

Appendix 1 : Literature Review and Recommendations on Teacher Incentives

While TaRL programs have been widely studied, evidence on the role of teacher incentives within these programs remains limited (Poverty Action Lab, 2023). Broader research on teacher incentives presents mixed findings; some programs enhance student performance, while others demonstrate little to no effect (World Bank, 2022). In some instances, incentives have had negative effects. For example, high-stakes testing and performance-based rewards can lead to unintended behaviors, such as "teaching to the test," manipulating test results, and reducing collaboration among teachers (Edwards & Roy, 2017).

John Hattie's *Visible Learning* provides further evidence on the variability of incentive effectiveness, synthesizing over 800 meta-analyses covering more than 80 million students to identify factors influencing educational achievement. Within this extensive review, Hattie examined the impact of various educational interventions, including financial incentives for teachers. While he found that financial incentives had a moderate effect on student achievement, their impact varied significantly across different contexts. This suggests that the effectiveness of teacher incentives depends not only on the financial reward itself but also on broader implementation strategies and local conditions.

Similarly, research from the National Bureau of Economic Research (NBER, 2019) in *Designing Effective Teacher Performance Pay Programs*, highlights the importance of both incentive design and the socio-cultural context in determining effectiveness. The report identifies key factors that influence effectiveness, including the types of performance measures used (e.g., student test scores, classroom observations, hybrid models), the nature of the rewards offered (e.g., financial bonuses, professional development opportunities, community recognition, hybrid models), and whether the incentives are structured for individuals or groups. It emphasizes the importance of aligning incentive structures with the socio-cultural context, highlighting that incentives are highly context dependent and models cannot be easily transferred from one setting to another..

If FAAQ decides to pursue teacher incentives, we recommend engaging in a thoughtful design process tailored to the local context:

- **Clarify Goals.** Before designing or implementing incentives, it is essential to clarify the goals they aim to achieve, assess whether alternative approaches might be more effective, and test assumptions. For example, if FAAQ seeks to improve the adoption of ENAd practices and enhance the quality of instruction, it could consider implementing school-wide or district-wide instructional rounds and professional learning communities (PLCs).

Research shows that these approaches significantly enhance teaching strategies and instructional quality, enabling teachers to collaboratively reflect upon their practices, receive and provide mutual support, and analyze student data and take collective responsibility for improving student learning. Research shows that these approaches significantly enhance teaching strategies and instructional quality by enabling teachers to collaboratively reflect on their practices and provide mutual support (Vescio, Ross, & Adams, 2008). Additionally, they foster strong cultures of learning among teachers and school principals, promoting continuous improvement at the school level, allowing educators to analyze student data together and take collective responsibility for enhancing student outcomes (City, Elmore, Fiarman, & Teitel, 2009).

- **Leverage Insights from Psychology and Behavioral Science.** Established research on motivation and behavior change could inform teacher incentive strategies and enhance FAAQ's understanding of program adoption and engagement data. For example, Self-Determination Theory emphasizes the importance of autonomy, competence, and relatedness in fostering motivation, highlighting the complex and sometimes unexpected interactions between intrinsic and extrinsic motivators. Additionally, frameworks such as Nudge Theory and the COM-B Model illustrate that subtle modifications in how choices are presented can profoundly influence behavior, with behavior emerging from the interaction of capability, opportunity, and motivation. Furthermore, Cognitive Dissonance Theory and the concept of social proof reveal that individuals experience discomfort when their beliefs and actions are misaligned, emphasizing the importance of aligning incentives (and communication about incentives) with teachers' core values and ensuring these values are reinforced within their specific social and cultural contexts.

- **View Teachers as Primary Users and Adopt a Strength-Based and Human-Centered Design Approach:** FAAQ could adopt a strength-based and human-centered design approach in all aspects of its program design, including its engagement with teachers and exploration of incentives. This approach could incorporate the following **methods**, with examples tailored to teacher incentives.

- **Contextual Inquiry and Ethnographic Research:** Gaining insights into teachers' experiences and environments.
- **System and School-Level Power Mapping:** Identifying the dynamics that influence decision-making within educational settings.
- **Empathy Mapping, Asset Mapping, and Journey Mapping:** Understanding teachers' perspectives, strengths, and the experiences they navigate.
- **Co-Design Opportunities:** Collaborating with teachers to create solutions that meet their needs.

By utilizing these methods, FAAQ can identify what truly matters to teachers, test assumptions, and appreciate the barriers and challenges they face, as well as the strengths that can be leveraged and built upon. Ultimately, this holistic understanding will enable FAAQ to better address teachers' needs and develop more effective programs.

- **Leverage FAAQ's Distinct Assets:** FAAQ might also explore the unique value it can provide. While other organizations may offer financial incentives or professional development, FAAQ has tremendous social capital that could be leveraged and may be more impactful than traditional incentives (e.g., Morat surprising a high-performing teacher or giving a concert to a school, investing in school or community development).

While these recommendations can inform the design of locally relevant and effective incentives, we strongly encourage FAAQ to consider them more broadly. These approaches could deepen FAAQ's understanding of the Colombian education context and its key stakeholders, providing FAAQ with valuable insights and an opportunity to build trust among key stakeholders. Such insights and trust could inform and support the expansion of the ENAD program and are crucial for establishing FAAQ as the go-to organization for adapting and scaling programs in Colombia.

Appendix 2:

2.1 Details on how to create pivot tables for analyses in Excel

Step 1. Pivot table in excel with # kids at each level per school, including # hours and average grade (please see appendix for detailed step-by-step guide)

Rows: by region (Tipo de sede: R/U), by school (Sede educativa)
Values:
Average of Número de horas hasta LM
Average of Número de horas entre LM y LF
Average of Número de horas TOTALES
Sum of LB Principiante L
Sum of LB Letra
Sum of LB Palabra
Sum of LB Párrafo
Sum of LB Cuento
Sum of LB Comprensión
Sum of LM Principiante L
Sum of LM Letra
Sum of LM Palabra
Sum of LM Párrafo
Sum of LM Cuento
Sum of LM Comprensión
Sum of LF Principiante L
Sum of LF Letra
Sum of LF Palabra
Sum of LF Párrafo
Sum of LF Cuento
Sum of LF Comprensión
Average of Grado (to get insight in the average grade per school)
Stdev of Grado
Min. of Grado
Max. of Grado
Optional: Filters: Grado: select 3,4,5 (if you want to select grades 3-5 only)

NB: make sure not to include the summary row in the pivot table

Step 2. Copy pivot table & calculate percentages per school
To rearrange and play with the data from the pivot table, copy the pivot table to a new tab with 'paste special > values'

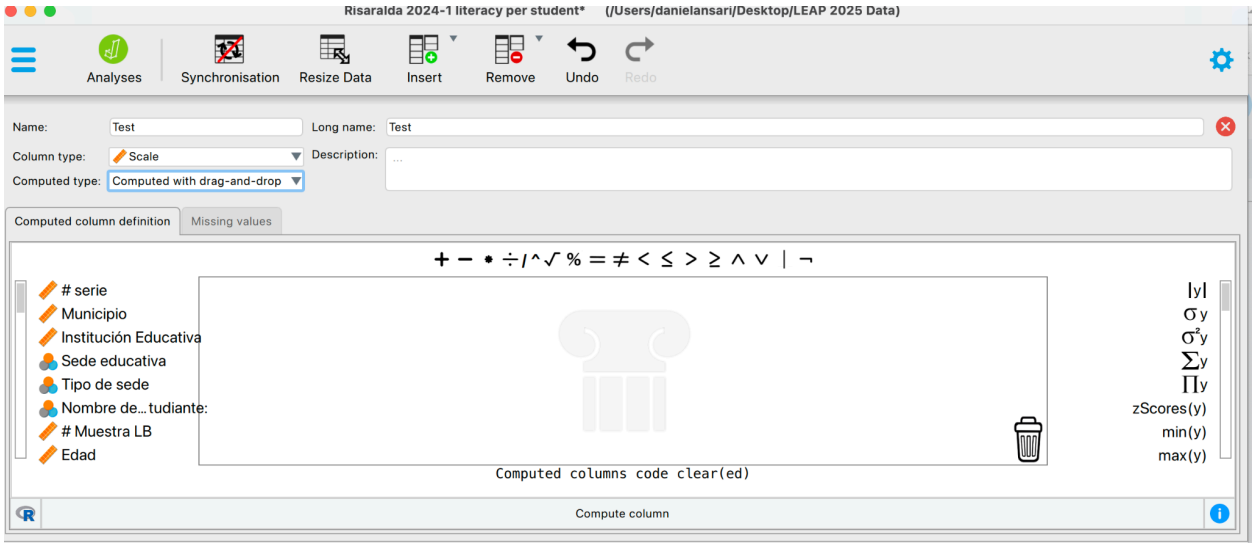
- Add a column 'Rural vs Urban'. Add 0 for each Rural school and 1 for each urban school.
- Calculate number of students at each level for each school at baseline, midline, endline. This allows you to select schools in your analyses who have enough students to calculate a meaningful percentage.
    - Formula =SUM(H44:M44)
    - When analyzing the 2024-1 Risaralda sample, we noticed that the rural schools all had <30 students (often <10). Because these low numbers

might bias the findings, we decided to use the average of all rural schools, which is displayed in the 'R' row in the pivot table. Because only 2 rural schools included the number of hours, we deleted the average number of hours across all rural schools from the table.

- Calculate % of students at each level for each school at baseline, midline, endline
    - at least level 3: Formula =SUM(K44:M44)/SUM(H44:M44)*100
    - at least level 4: Formula =SUM(L44:M44)/SUM(H44:M44)*100
    - highest level: Formula =M44/SUM(H44:M44)*100
    - Give the columns a name, for example B345pct, where B = baseline, pct = percentage of students, and 345 = means achievement level 3-5.
- Delete rows that are not used. For the 2024-1 Risaralda sample, that would be row **U** (which shows the average of all urban schools), row **Grand total** (which shows the average of *all* schools), and all rows with individual rural schools.
- Copy data to JASP by selecting the table including the headers and pasting it into JASP.

2.2 How to calculate how many students are at a given level at 2 or more time points.

In order to calculate how many students are at a given level for 2 or more time points we first need to create a new variable. In JASP go to 'Edit Data' then select the last variable and press the green plus sign. This will create a new variable, give it a name and select 'Compute with drag-and-drop' from the menu entitled 'Computed type':



Now you can specify how you would like to compute the variable. For example, if you wanted to have the variable show whether a student was at level 2 of the ASER Literacy Test for both midline and endline, you would compute Literacy Midline + Literacy Endline = 2. This would then mean that your new variable would output a 1 for every student that was at level 2 for both midline and endline. You can then use the descriptive function to calculate how many students fulfill this computed criteria.

References

Abdul Latif Jameel Poverty Action Lab (J-PAL). (2023). Teaching and incentives: Substitutes or complements? Retrieved from https://poverty-action.org/sites/default/files/2023-07/Teaching-and-Incentives-Substitutes-or-Complements%3F.pdf

Brouwer, N., Besselink, M., & Oosterheert, I. (2017). The power of video feedback with structured viewing guides. *Educational Technology Research and Development*, 65(5), 1391-1411. https://doi.org/10.1007/s11423-017-9500-0

City, E. A., Elmore, R. F., Fiarman, S. E., & Teitel, L. (2009). *Instructional Rounds in Education: A Network Approach to Improving Teaching and Learning.* Harvard Education Press.
Edwards, M., & Roy, S. (2017). Perverse incentives in education policy. *Science Policy Review*. Retrieved from https://svpow.com

Danielson, C. (2013). *The Framework for Teaching Evaluation Instrument*. The Danielson Group. Retrieved from https://danielsongroup.org/framework/

Data Wise Project. (n.d.). *Data Wise: A Step-by-Step Guide to Using Assessment Results to Improve Teaching and Learning*. Harvard Graduate School of Education. Retrieved April 7, 2025, from https://datawise.gse.harvard.edu/

Hattie, J. (2017). *Visible Learning: A Synthesis of Over 800 Meta-Analyses Relating to Achievement.* Routledge.

Jacobs Foundation. (n.d.). *Education Evidence Labs (EdLabs)*. Retrieved April 1, 2025, from https://jacobsfoundation.org/education-evidence-labs-edlabs/

Kane, T. J., Taylor, E. S., Tyler, J. H., & Wooten, A. (2011). Identifying effective classroom practices using student achievement data. *Journal of Human Resources*, 46(3), 587-615. https://doi.org/10.3368/jhr.46.3.587

Klette, K. (2023). Classroom observation as a means of understanding teaching quality: Towards a shared language of teaching? *Journal of Curriculum Studies*, 55(1), 49-62. https://doi.org/10.1080/00220272.2023.2172360

National Bureau of Economic Research (NBER). *Designing Effective Teacher Performance Pay Programs*. 2019.

Ridge, N. & Lavigne, A. (2020). The impact of peer observation on teaching quality: A systematic review. *Teaching and Teacher Education*, 89, 102963. https://doi.org/10.1016/j.tate.2019.102963

Sherin, M. G., & van Es, E. A. (2009). Using video to support teachers' ability to interpret classroom interactions. *Journal of Technology and Teacher Education*, 17(2), 119-144. https://www.learntechlib.org/p/29505/

Vescio, V., Ross, D., & Adams, A. (2008). A review of the research on the impact of professional learning communities on teaching practice and student learning. *Teaching and Teacher Education,* 24(1), 80-91.

World Bank. (2022). School resources, better teacher incentives, or both to improve student learning. Retrieved from https://www.povertyactionlab.org/evaluation/more-school-resources-better-teacher-incentives-or-both-improve-student-learning-0?lang=es

World Bank. (2018). *TEACH: Classroom Observation Tool Manual*. Retrieved from https://documents.worldbank.org/en/publication/documents-reports/documentdetail/465231542639159875/teach-classroom-observation-tool-manual