

2023 LEAP CHALLENGE

Exploring the Effects of Play-Based Learning in Early Childhood Education in Lebanon and Jordan

Project Host:



Fellows:

Jazib Zahir, Social Entrepreneur Fellow

Rene Kizilcec, Research Fellow

Sergio Medina, Social Entrepreneur Fellow

Sho Tsuji, Research Fellow

Spring 2024

SUMMARY

A great benefit of the LEAP program is that it allows host organizations and several experts in their respective fields to rally around a collective question: how can the host organization supercharge their path to have rigorous evidence of the effects of their program?

Over the course of a team sprint this Spring, we set out to address that question for the Seenaryo Playkit - a teacher training delivered alongside an app-based learning tool, that provides easy access to hundreds of play-based learning activities that are designed to build social and emotional learning (SEL) and life skills for children ages 3-8 in Jordan and Lebanon. The Playkit is a program of Seenaryo who operate in Jordan and Lebanon and to date has reached over 400 schools, serving over 89,000 children.

Seenaryo has been operating in the field in Lebanon since 2015 and Jordan since 2018 and is one of few organizations championing play-based learning as an essential component of early childhood instruction in the region.

As we began to answer this collective question of “How Might We Supercharge the Path to Rigorous Evidence of the Program Effects,” several sub-questions emerged:

- **EVALUATING IMPACT:** How does Seenaryo ready themselves as an organization for a future experimental study of the effects of the Playkit?
- **CHILD EFFECTS:** What trusted/tested/rigorous methods might Seenaryo use to observe the effects of the Playkit on children in the classroom?
- **GREAT DATA:** What are some of the best practices of capturing and analyzing Playkit data and how does Seenaryo build that capability?
- **APP DATA:** How can Seenaryo get the most information from the app and tech-based environment (surveys) while optimizing automation?

As we explored these many questions in the sprint, a focus to our work emerged when Seenaryo shared what we began to refer to as their North Star Metrics -- the metrics that they theorize would demonstrate positive intended effects on their students as they were designed.

The impetus to build a stronger evidence base is guided by Seenaryo's interest in building a better and more effective product, demonstrating that the product is having intended positive effects on children's learning and outcomes, and build a case to drive greater investment in Seenaryo so that they may scale within Jordan and Lebanon, and set the stage for additional regional scaling.

EVALUATING IMPACT

Seenaryo is ready to take the steps to prepare for an impact evaluation that matches where they are at as an organization. With a program that faces a wide variety of operating contexts within their region, both within and between countries, we found it useful to reassess and simplify Seenaryo's impact claim.

Seenaryo took a moment to hone their North Star Metrics that they theorize will provide evidence of the positive change that occurs in children when play-based learning activities are incorporated into the delivery of the as-usual curriculum. With these refreshed metrics, they have the basis for an updated Theory of Change which serves two key purposes of (1) clearly communicating what is the intended impact of the product they designed and (2) it serves as the basis of a future impact evaluation which is designed based on their theory of change and how the impact evaluation can be constructed to evaluate that question.

The designers of the intervention play a vital role in education content and curriculum development. The design team at Seenaryo is deeply rooted in arts education and developed the Playkit as a purposeful intervention to bring all the benefits and intended positive effects of games, songs and stories to children in the Middle East who face a unique set of socioeconomic and geopolitical contextual challenges.

Before we can map where Seenaryo can go in their impact research journey and answer our primary question of “how can the host organization supercharge their path to have rigorous evidence of the effects of their program?”, we first have to assess where Seenaryo falls on the [NESTA Standards of Evidence](#) continuum.

NESTA is a prominent social innovation agency/foundation for social good in the UK and developed a widely-used model on how to chart organizations based on their progression along a continuum of developing rigorous standards of evidence of how a program’s interventions perform.

NESTA has 5 standards along their continuum:

- Level 1:** You can describe what you do and why it matters, logically, coherently, and convincingly
- Level 2:** You capture data that shows positive change, but you cannot confirm you caused this
- Level 3:** You can demonstrate causality using a control or comparison group
- Level 4:** You have one + independent replication evaluations that confirms these conclusions
- Level 5:** You have manuals, systems and procedures to ensure consistent replication and intended impact

Figure 1

The Nesta Standards of Evidence

The objective of developing Standards of Evidence is to help us know how confident we can be in the evidence provided to show that an intervention is having a positive impact.



Source: Puttick, R. and Ludlow, J. (2012) 'Standards of Evidence for Impact Investing.' London: Nesta.

Based on a review of where Seenaryo sits on the continuum, we agreed that they are at Level 2 and they want to chart paths to reach level 3 in the near-term and to reach level 4 in the medium-term.

The foundational component of these paths/roadmaps to move from Level 2 to Level 3 (in the medium term, 1-2 years), and then level 4 (in long term, 2-5 years), begins with mapping the intended impacts of Seenaryo's designed activities. The intended impacts are supported by their North Star Metrics.

Currently, Seenaryo relies on three sets of data tracks to the scale and impact of their work: activities, usage data, and user feedback. Seenaryo tracks all teacher training (activities), all app downloads, all usage within the app (usage data), and the dissemination and analysis of surveys designed to capture the intended effects of the Playkit and understand their user impressions (user feedback).

Moving from Level 2 to Level 3 along NESTA's Standards of Evidence will require some retooling of their data tracking methodologies. ***As Seenaryo moves to higher levels of research rigor, they will shift from an impression/survey-based approach to a more student and observation-based impact evaluation that aims to demonstrate causality (e.g. intervention x shows an intended effect in y) using a control or comparison group.*** The next-level up is having an independent replication evaluation, and then agency-wide systems in place to maintain those effects and know from data that is what shows - landing them at an aspirational Level 5.

It is easier said than done for a nonprofit organization such as Seenaryo juggling operations in two countries across hundreds of schools with thousands of affected students. Moving from Level 2 to Level 3 is a time-intensive and expensive effort for any organization. Funding doesn't often support the high costs of impact/evaluation studies, and when they do - they can catapult organizations to scale when there is evidence of positive effects.

By defining the intended effects of their intervention, as a first step, Seenaryo is readying the Playkit for an impact evaluation that can move them from Level 2 to Level 3 and further prepare themselves for moving toward Level 4 and perhaps Level 5.

Seenaryo’s North Star Metrics

Table 1

INTENDED IMPACT / OUTCOMES	NORTH STAR METRICS
<ol style="list-style-type: none"> 1. Improved Communication Skills 2. An increase in cooperation 3. An increase in emotional resilience 4. An increase in emotional self-regulation 5. Improved fine motor skills 6. Improved Decision-Making Skills 7. Improved Problem-Solving Skills 	<ul style="list-style-type: none"> ● Children who are introduced to and instructed with Playkit’s play-based learning activities show positive effects on Skills 1 - 7. ● A plan would need to be developed to document changes in children provided with Playkit, beginning with building capacity internally at Seenaryo and moving to partner with a research agency to conduct the study
<ol style="list-style-type: none"> 8. Improved Learning Environment 	<ul style="list-style-type: none"> ● Seenaryo identified this impact as a guiding goal. Two possible complementary metrics could be: Teacher Retention and Student Attendance Over Time and Drop-Out Rate ● Where data doesn’t exist - there are ways to establish baselines
<ol style="list-style-type: none"> 9. Increased Teacher Engagement w/ the Playkit 	<ul style="list-style-type: none"> ● Seenaryo would need to define what engagement entails. E.g. - engagement could be usage of learning activities on the app and a metric derived from the positive Δ in user accessing of the various activities
<ol style="list-style-type: none"> 10. Improvements in Literacy and Numeracy 	<ul style="list-style-type: none"> ● Test scores or evaluations ● Seenaryo noted that there are no national test scores for the 3-8 age group in Jordan and Lebanon - so if they are to look at this as a North Star Metric, measure will need to be put in place to establish baselines and pre-post comparisons in some way
<ol style="list-style-type: none"> 11. Improvements in well-being* 	<ul style="list-style-type: none"> ● Inclusion has been an intended effect that has been designed for. If Seenaryo were track this effect, they would need to design a part of the study to observe and document changes in “inclusion” with target research group ● Other indicators of improvement in well-being in the classroom include: Positive attitudes towards outgroups, Pro social behaviors, and Reduction in behavior challenges. In each of these categories, to build a study to document the effects of Playkit on these aspects, Seenaryo would need to very carefully design the study to isolate the intended effect.

A useful exercise here to gain more insights into the workings of the Playkit, would be to workshop all activities in the Playkit app and show how each and every activity maps to and supports one of the 7 intended impacts areas (or more) - and in that process - decide on which of the 11 indicators total are your highest priority and build study strategy based on the prioritized outcomes.

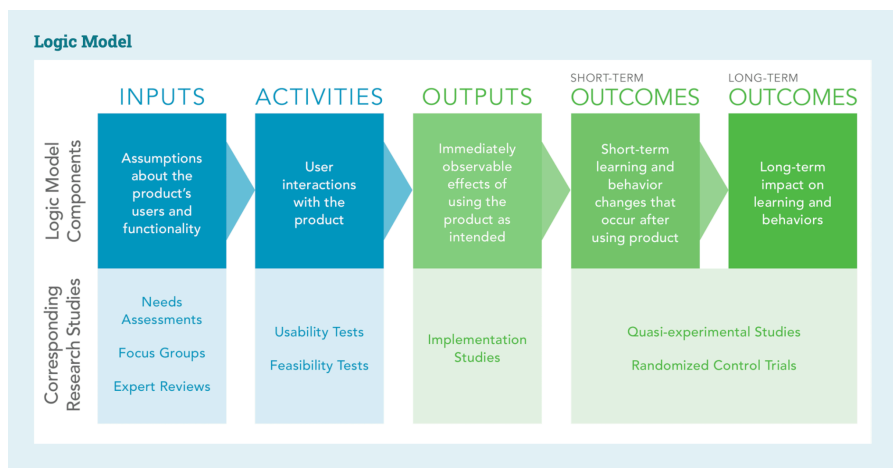
Refreshed Theory of Change

Table 2

ACTIVITY	SHORT-TERM EFFECTS	LONG-TERM EFFECTS
Children aged 3-8 in Jordan and Lebanon are introduced to and consistently supported to participate in play-based learning activities provided in the Playkit to teachers and...those activities are delivered and integrated into the widespread everyday instruction	In classrooms where the Playkit is implemented, we will see immediately observable positive effects in <ul style="list-style-type: none"> * Communication * Cooperation * Emotional Resilience * Emotional Self-Regulation * Improved Motor Skills 	Children who receive play-based learning have <ul style="list-style-type: none"> * Improved Decision-Making Skills * Improved Problem Solving Skills And those skills translate into <ul style="list-style-type: none"> * Increased lifelong earnings * Increased employment * Improved Well-Being

WestEd Logic Model

Figure 2



Refreshed Seenaryo Logic Model:

Table 3

	Inputs	Activities	Outputs	Short-Term Outcomes	Long-Term Outcomes
COMPONENTS:	Assumptions about the product's users and functionality	User Interactions with the Product	Immediately observable effects of using the product as intended	Short-term learning and behavior changes that occur after using product	Long-term impact on learning and behaviors
RESEARCH STUDIES:	<ul style="list-style-type: none"> *community needs assessments to design intervention *meets with beneficiaries (focus groups) to understand product in place * expert reviews: gather feedback from relevant contributors to improve product 	<ul style="list-style-type: none"> * Product usability Tests * Feasibility studies for product and program functioning at scale 	*Implementation Studies	<ul style="list-style-type: none"> * Quasi-experimental studies * Randomized Control Trials 	

Seenaryo Advancing to Levels 3 and 4

At the moment, Seenaryo has completed multiple community needs assessments, has run a high number of focus groups and they are engaged with a number of expert reviews of their interventions. Through tech teams and consultants, they've conducted a number of product usability tests and have assessed the feasibility of in-country and regional expansion. As well, they have assessed barriers to scaling to better help them understand how to implement the Playkit on a wider scale.

In the Playkit's journey to prepare for an experimental study (where a control group is compared to a research group), it is helpful to look to two organizations who have done an exceptional job at documenting the effects of their early years play-based

interventions, and they are the International Rescue Committee and Sesame Street, and Lively Minds. While there are many differences between these organizations and Seenaryo, their region, and their specific interventions when comparing them to the Playkit, there are enough similarities to look to them as examples of programs that show the positive effects of play-based learning activities for children aged 3-8 in developing countries.

IRC / Sesame Street

In the past 5 years, the MacArthur Foundation funded the International Rescue Committee (IRC) and Sesame Street \$100,000,000 USD to address the grand social challenge of bringing safer and more stable educational options to children in emergency and conflict settings.

The MacArthur Foundation funded the partnership, titled, “*Ahlan Simsim*” - a blended variation of IRC’s and Sesame Street’s work - adapted to the local language and cultural context.

Several years into implementation, they conducted an RCT to evaluate the impacts of showing a set number of episodes of Ahlan Simsim (play-based and mass-media-delivered curriculum) to a set number of schools in Lebanon and comparing those effects to the same set number of schools who did not receive the intervention in Lebanon. That allows for a control group comparison effect.

The findings of the study were very positive and promising. Partnering with the [Global TIES for Children at NYU](#), Ahlan Simsim set out to evaluate the effects of their program as designed and as intended.

The intervention design was led by Sesame Street (and advised by IRC). Sesame Street is a U.S.-based and decades-long leader in early childhood education and they, and field practitioners from IRC, designed Ahlan Simsen shows and characters with the intention to build SEL in children.

The major finding of the study was that children shown a set number of episodes of Ahlan Simsim programming were both (1) better able to recognize, label, and

categorize a range of targeted emotions, and (2) more likely to use “I’m feeling X emotion and I can use breathing exercises to manage that emotion.”

While these findings might not seem earth-shattering on the surface, they break very important ground in documenting, through rigorous social scientific inquiry and research adhering to widely-accepted research practices, new instances of positive causal effects on student learning outcomes and their learning that acts as a launching pad for positive outcomes as an adult.

The IRC study builds upon other research studies that have shown a causal positive link between increases in SEL and increases in literacy and numeracy, prosocial behaviors, and inclusion of outgroups - and these skills are also shown to result in higher socioeconomic outcomes as desirable long-term outcomes.

As Seenaryo develops their own impact evaluation plan, it can be helpful to focus heavily on the design of the playkit and delve even deeper into each of their activities, why it was designed the way it was, what are the educated theories the team has about what skills each activity introduces and reinforces, and what are the short and long term effects of the 7 core skills, for example.

Once the designed effects of each activity are all mapped out - those effects can be measured with evaluation tools - many of which are described throughout this report.

Lively Minds

In the past several years, the organization Lively Minds has achieved incredible outcomes for preschool children in Ghana and Uganda using play-based learning. They partnered with [M.I.T. 's Jameel Abdul Latif Poverty Action Lab \(J-PAL\)](#) to design and implement an RCT to document the effects of their curriculum.

The results of the study are still being finalized and will be published shortly, but a summary released recently shared that the results are indeed significant.

The study found that children provided with Lively Minds’ curriculum showed positive effects on cognition, school readiness, and increases in pre-numeracy, pre-literacy,

and fine motor skills. The study also observed positive effects on children's social-emotional development coupled with health improvements. In participating in the Lively Minds program, parents showed increases in self-esteem and greater involvement with their children's teachers.

While there are many differences between Seenaryo and Lively Minds, the findings of their study are affirming of the expansion of play-based learning. Moreover, the results of their study are good for the field and good for the claim that investing in play-based learning can have profound and enduring impacts on children's development.

Seenaryo has their own unique combination of SEL and life skills they are targeting through the Playkit - and their ultimate validation study will evaluate their specific intended effect. But taking cues from Lively Minds (particularly when the study is published), can save Seenaryo some time and effort, and can reference the study to show that indeed, investing in play-based learning is a well-documented, cost-effective and proven way to positively impact children's development.

GREAT DATA

Build organizational capacity to increase and smooth data capture and analysis, and reduce staff data processing

Seenaryo is at a pivotal moment from the perspective of data capture and analysis for impact evaluation and internal prioritization. The organization's reach has been growing quickly, with many new teachers getting trained every month, each reaching several dozen students year after year. In most of its reports, Seenaryo has relied on responses collected from teachers via Google surveys at various points in time. These data are currently aggregated and analyzed in Google spreadsheets. Several of our conversations during the LEAP period have focused on building organizational capacity to transition to a robust data infrastructure that supports Seenaryo's reporting needs as it keeps growing. This infrastructure should enhance the efficiency of data processing to reduce the amount of manual labor.

We suggest three-part organization of the data strategy to facilitate the planning process for Seenaryo. The first part is to evaluate and update the specific measures

that are collected (e.g., what survey questions to ask and what response scales to use). Here it is important, as described in more detail below, to align measures with key objectives. The second part is how the data is represented in the data lake where different data sources are stored to facilitate analysis. Here it is important to bring on external help to set this up in a way that enables you to get the most out of the data that is collected. A clear vision of the insights you require (e.g., pre-post comparisons for individual questions) will help set up the data lake in the right way to facilitate these analyses. The third part is the analysis strategy which will depend on the type of data that is being analyzed (e.g., counts of clicks or video plays, ordinal survey responses on agree-disagree scales). Selecting the appropriate analysis strategy given the data and question at hand is an important step towards extracting robust evidence to include in reports and inform internal decisions. Here it is once again important to bring in external help to ensure that the analysis strategy is appropriate and it is clear how to interpret statistical findings.

We started the process of reviewing existing data collection, storage, and analysis processes, by mapping out the “northstar” goals of Seenaryo (at least with respect to the Playkit, which was the focus of our LEAP collaboration). This yielded a list of six goals which we considered as targets for measurement. We considered not just the data collection methods that Seenaryo is using right now, but also potential new methods that could be adopted. In Table 4, we explored alternative methods for measuring each northstar goal to understand if and how the current data collection approach needs to change. We consider surveys (primarily of teachers, but possibly also of school leaders and parents), in-class audio recording as a passive and objective way to measure what goes on in class, user log data from the Playkit App to measure what resources teachers go to in the app, and large-scale records provided by the government or non-governmental organizations that measure macro-level trends in and across schools that Seenaryo operates in (e.g., teacher turn-over, student attendance and performance).

We consider this activity of aligning goals with measurement approaches and also the specific measures that are being collected (e.g., what questions are asked on the survey) to be valuable to Seenaryo going forward, and worth repeating on a regular basis to adapt their data strategy to any changes in their northstar goals. In the remainder of this section, we review potential innovations for Seenaryo’s data

collection efforts and finally an approach for consolidating data to derive deeper insights by combining data from different sources.

Recommendation: *Intentionally align the data strategy with Seenaryo’s northstar goals and regularly monitor this alignment as goals evolve, and probe the effectiveness of data sources to provide meaningful insight on progress towards specific goals.*

Table 4. Mapping the Northstar Goals of Seenaryo’s PlayKit to Potential Data Sources.

Expected Impacts of Seenaryo’s PlayKit	Survey	In-class Audio Recording	PlayKit App Log	Large-scale Records
A transformation in the school environment	X	X		X
Thousands of teachers excited to be part of a community / increase in teacher retention	X			X
Test scores / academic achievement	X			X
Children are happier / increase in their wellbeing	X	X		
N schools have reported a reduction in behavioral challenges	X			X
Increase in the development of the seven life skills that the Playkit targets	X			

Survey Data Collection

Surveys are one of the primary data sources of Seenaryo’s evaluation efforts at this time. In particular, Seenaryo surveys teachers before and after their professional development training, as well as a few weeks later to check in once more. In this section, we examine the content of the survey and ways to analyze the data. Here we focus on the data collection method itself which has important implications for the type of data that can be collected and how it can be analyzed. Seenaryo currently uses Google Forms to collect data, which is free, simple to use, and well-integrated with the rest of the Google Drive suite of applications. However, Google Forms have a number of limitations which may not be apparent until they are compared to alternative

technologies for survey data collection. We drew a comparison to both LimeSurvey and Qualtrics, which are not free and require learning a new tool, but can provide useful new features. This is a tradeoff worth careful consideration, as transitioning to a new survey platform incurs a significant logistical cost. We describe a few of the features of more advanced survey tools like Qualtrics or LimeSurvey here to highlight their utility:

- Question types: Advanced survey tools provide a larger selection of question types that may be easier to respond to: sliders, images as responses, matrix-style rating questions, and more.
- Conditional questions: Skip questions that are not relevant to some respondents. Show different sets of questions based on an earlier response, or prior knowledge of the respondent (e.g., what school they are in).
- Translation: Advanced survey tools help with translation and store the questions/responses in several languages. That way a respondent can select their preferred language to answer the survey.
- Response encoding: The responses are encoded as both verbal (e.g., strongly agree) and numeric (e.g., 7) data. This saves time when analyzing data and reduces the likelihood of human error.
- Respondent tracking and reminders: Provide the tool with a list of emails that you expect a response from and it will track who has not responded yet and send reminders to increase the response rate.
- Metadata: Advanced survey tools allow for metadata to be recorded along with the responses to connect data points. For example, a response can have metadata such as a unique ID for the school and for the teacher in order to match responses over time (or connect survey data to other data sources) and analyze individual trends.
- Analysis capabilities: Advanced survey tools provide analysis capabilities that simplify data visualization, descriptive and statistical data analysis, and even text analysis (e.g, word clouds or topic modeling of open-ended responses).

Recommendation: Consider adopting a more sophisticated survey tool to gain access to features that can make data collection and analysis more efficient and versatile.

Classroom Audio Data Collection

To gain a better understanding of how teachers integrate the materials from the PlayKit and the techniques learned during the teacher training in their classroom practice, we

considered classroom audio recording as a data source. While in-person classroom observation can provide richer and more contextualized data, it is harder to scale and might be subject to more bias. Classroom audio recordings can be used to extract a number of useful measures for Seenaryo, including the quantity of speech, the distribution of speech by different talkers, and conversational turns. Moreover, though this depends on the quality of microphones and background noise, classroom recordings can also be used to provide feedback on the content of what is being said (e.g., how well the teacher engages with the questions/comments from students, how much the conversations align with the curriculum). Overall, this data source can provide quantitative insights into classroom dynamics and serve as a pre- and post-measure to assess the impact of introducing Seenaryo on classroom practices.

We explored three alternatives for integrating classroom audio recordings into Seenaryo's data collection efforts. The first option is to embed audio recording capability into the Playkit app, which may be cost-effective in the long run, but require additional software development and processing. The up-front cost may be too high especially given the availability of established alternatives in the market. One such alternative is LENA (<https://www.lena.org/>), a portable device to record classroom talk, which is widely used in child development research. It comes with software that has in-built features relevant to data protection. However, the cost of purchasing the devices (wearable audio recorders that look like vests) and analysis software licenses may be too high without a clear case for the benefits of this approach.

Finally, we considered a relatively new product on the market that is used by teachers across the United States: TeachFX (<https://teachfx.com/>). It is a mobile and desktop application that uses a teacher's or the school's own device for recording (e.g., the teacher's smartphone). The application works in different languages and has a dashboard interface that provides AI-based feedback on student and teacher dynamics. The company offers a free trial which could be used in a pilot to evaluate the quality and utility of this data source.

Recommendation: *Pilot a classroom recording application to explore this new data source as a way of understanding classroom dynamics with Playkit, and perhaps even to show pre-post differences in classroom activity after adopting Playkit.*

Behavioral App Data Collection

Teachers use the Playkit app to browse and review course materials. There is a significant opportunity here to set up a data collection pipeline to understand who is looking at what, and perhaps even ask teachers to report what materials they have used, how satisfied they are with the experience of using it. The Playkit app presents an ongoing interactive relationship between Seenaryo and the teachers. Other sections in this report provided more detail on ways to leverage this data source. Here we highlight the importance of aligning data collection from the app with specific questions that are of interest to Seenaryo to ensure that the right data is getting collected to answer those questions. This may involve asking the developer to add in a few feedback elements, such as a check-mark feature for teachers to report that they used a piece of content, or a how-did-it-go feature to report back on the classroom experience using a piece of content. This type of information can be useful to Seenaryo to evaluate and review their own content, and also for other teachers who might be interested in what content is popular and easy to implement well with students.

Recommendation: *Leverage behavioral data from the Playkit app in line with key questions and integrate feedback mechanisms that collect additional useful information for Seenaryo and potentially other teachers.*

The Playkit app is about to be relaunched after an overhaul of the user interface and system architecture. The system does not require users to log in to reduce barriers to access. However, one of the downsides of not requiring any authentication is that it is difficult to track users over time. This can be useful for three key reasons: (1) to understand who the user is (what school if they are a teacher), (2) to understand what materials individual teachers engage over a longer period of time, and (3) to connect data from the app with data collected from surveys and other sources.

Recommendation: *Add persistent identifiers for teachers in the PlayKit app to track the behavior of individual users over time and connect app usage data to other data sources such as surveys and audio recordings.*

Large-scale Record Collection

Seenaryo's northstar goals include a transformation of school and classroom culture that has implications on teacher retention and student standardized test scores. It may be feasible to track some of these macro-level effects of using the PlayKit using large-scale records collected by the government (e.g., the Ministry of Education and/or Labor) or non-governmental organizations (e.g., the Queen Rania Foundation). Impacting any such macro-level measures is a stretch goal, but monitoring the availability and trends in whatever data is available can be useful to Seenaryo nonetheless. For example, it can inform the targeted outreach to different schools to recruit teachers to use the PlayKit.

Recommendation: *Identify and track large-scale records of school performance, student performance in those schools, and teacher retention to look for evidence of temporal trends related to the adoption and increased use of the PlayKit, or to inform targeted outreach to schools and teachers.*

Consolidating Multiple Data Sources

It is evident that Seenaryo is at a pivotal moment for updating its data collection and storage infrastructure to gain better insights from the data that they are gathering. Currently, the data is stored in spreadsheets on Google Drive, which has a number of limitations that Seenaryo is aware of. To address these shortcomings, they are keen to adopt Qlik (<https://www.qlik.com/us>), which is a data storage solution with analysis capabilities. It is certainly not the only product on the market that can play this role. If implemented correctly, Qlik will enable Seenaryo to create a data lake: a place for all datasets to be stored alongside each other in ways that facilitate analyses across datasets. Moreover, it can reduce manual labor and the introduction of human error by automating processes such as data cleaning and transformation (e.g., transforming agree-disagree scales into numeric values). Qlik supports a large number of data sources including Google Forms, but also a number of other survey tools, including Qualtrics. It also provides tools for data visualization and analysis that can increase the efficiency of generating insights and support Seenaryo's data communication. The challenges for Seenaryo will be to use the features that Qlik offers. This will require setting up data that is already collected so that it can be analyzed; developing routines

that can run automatically to ingest new data, clean it, and generate insights; and preparing a data infrastructure that is expandable to include new data sources such as those noted above.

Recommendation: *Work with a partner to set up Qlik in such a way that leverages its features for the analysis of already collected data and provides a foundation for integrating and connecting new data sources to gain insights.*

Build organizational capacity to interpret and present data

Data analysis can entail description of results, data visualization, and statistical analysis. How to best go about these different parts depends on the type of data as well as the goals of the analysis and audience to be reached. For instance, a general audience might be better served by a single accessible visual or descriptive presentation, while a funder that looks for the quality of evidence might expect more comprehensive figures and a robust statistical analysis of the data. The Seenaryo team has shown impressive capacity to organize and describe the data. In this section, we add suggestions for two specific areas: first, data visualization to help both the organization as well as potential audiences to get a quick and easy-to-grasp overview of outcomes, and second, statistical data analysis as an additional tool to enhance the descriptive presentation of evidence.

The data and previous workflow

The Seenaryo team has carefully collected questionnaire data from teachers across the schools they are servicing. These questionnaire data have often been collected before and (immediately or a few weeks) after teachers participated in the Seenaryo trainings, and can therefore provide valuable insights into changes. Teachers were asked to rate to what extent they agree to statements capturing a wide range of their experiences (e.g., “I believe children learn best sitting behind a desk” on a scale of "Strongly Disagree", "Disagree", "Neither agree nor disagree", "Agree", "Strongly Agree"). These data points have been organized and stored in Google spreadsheets. Seenaryo has

Figure 3



more teachers understand how to **create an inclusive learning environment** (from **86%** to **98%**)

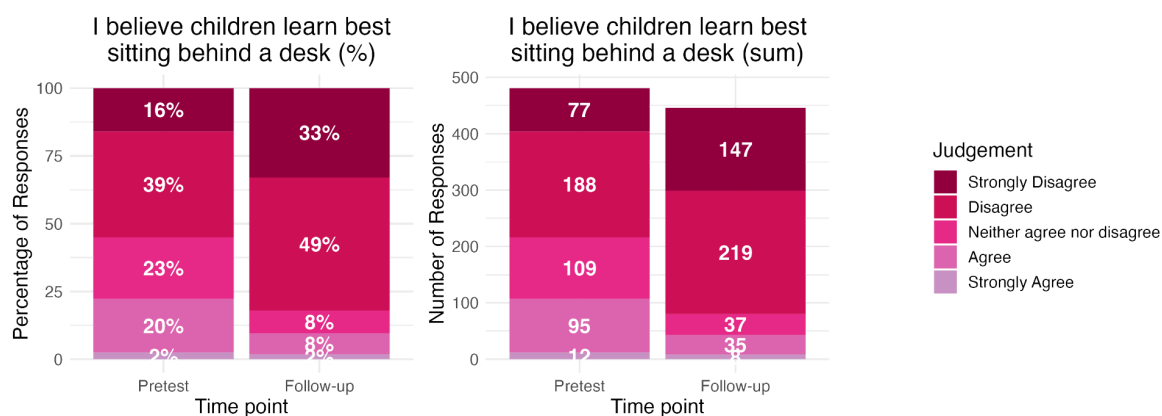
created automated functions to summarize and describe this

data, such that percentages of each response category and each question item are listed. From these tables, the team picked up noteworthy values by visual inspection - for instance, items for which the percentage of “Strongly Agree” changed quite a bit between pre- and post-assessment, and incorporated such values into reports (see illustration to the left). This approach makes it sufficiently possible to pick out remarkable changes, and can lead to impactful visualizations. We do, however, suggest to also incorporate elements of more systematic data visualization, to facilitate data interpretation both for the team and data-focused outside parties.

Data visualization

The data collected are called *ordinal data* in statistical lingo. Such data have several ordered categories, where the distances between the categories are not necessarily known. This means that we can not simply calculate an “average response” for a given question. So for the question “I believe children learn best sitting behind a desk”, it is not as simple as looking at the data and concluding “x% of teachers agree” but the outcome data show: “x% of teachers strongly agree, x% agree, x% neither agree nor disagree, x% disagree, x% strongly disagree”. In addition, if we consider that we often have pre- and post-training data, we would need to look at these five levels of responses in both pre- and follow-up-data sets to see whether there are any significant improvements. No matter how we later choose to combine and summarize the data (see next section), it will make sense to choose a strong visualization to get an impression of the pattern of responses and potential changes pre- and post-training. A powerful way to do so for ordinal data is a stacked bar graph (figure below).

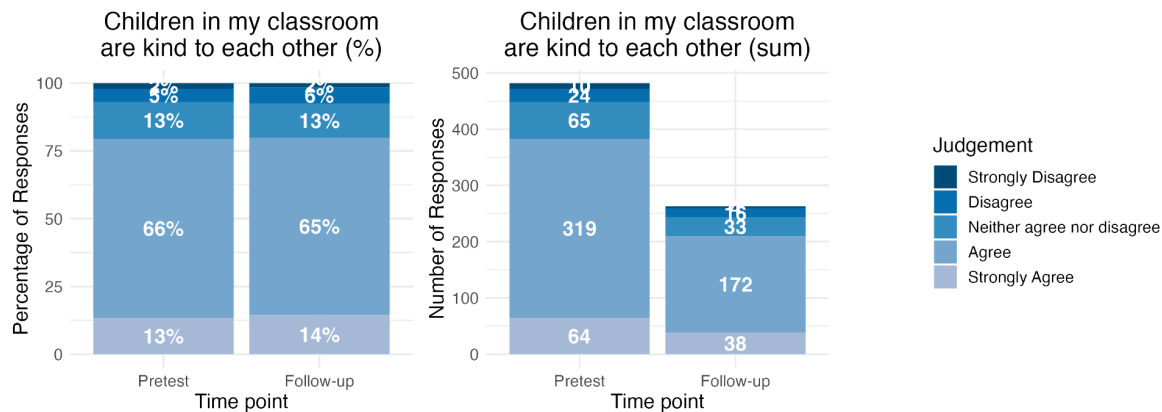
Figure 4



If such a stacked bar graph were generated for each survey item, it would allow the

Seenaryo team to get an easily interpretable overview of their data, and guide them towards which items warrant further exploration. Such graphs can be created easily within Google spreadsheets, the software solution Seenaryo is familiar with. Another consideration is whether to plot the percentage of responses (as plotted to the left) or the total number of responses (as plotted to the right). Percentages give us information about the relative change between pre- and post-test, and the fact that both estimates add up to 100% makes comparisons interpretable. This is therefore a correct representation of the data that allows us to compare different categories. One piece of information that we lose when converting numbers into percentages, however, are the absolute values. As can be seen on the right side of the above graph, there are less values in the follow-up-test than in the pre-test, meaning that less teachers filled in the questionnaires in the second test. In this case, the difference in numbers is not alarming, but it is still recommendable to also check the absolute numbers. For instance, consider the example below for the question “*Children in my classroom are kind to each other*”: Here, we can see, based on the total number of responses on the right side of the graph, that we have a much lower response rate in the follow-up-test compared to pre-test. When looking at percentages (left side), we would, for instance, think that the number of teachers that respond “Agree” is similar between pre- (66%) and follow-up- (65%) test, although the actual number of responses is much lower (319 vs. 172). Thus, based on these two different representations, we would not necessarily reach the same conclusion. Since it is hard to guess whether the responses of those teachers that did not respond would show similar tendencies, it seems fair to say that data for which the number of pre- and follow-up-responses starkly differ need to be interpreted with caution. Thus, a good workflow is to first check the absolute numbers (this can just be an overall count of responses per questionnaire item, without the breakdown into the different ratings), before looking into the percentage view.

Figure 5



Take-aways:

- Stacked bar graphs are a good way to visualize ordinal data, either for internal purposes or for data-savvy external audiences.
- It can be important to look at both absolute and relative numbers to understand the data.

Data presentation and interpretation

We have seen above how we can visually present ordinal data. So how do we talk about them? One good way, already chosen by Seenaryo, is to curate responses and present them to an audience. Looking back at Figure 4, it represents the results of an ordinal questionnaire item as *one* response. In the case of our running example of the question item “*I believe children learn best sitting behind a desk*”, we could sum up the percentages of the two response categories “Strongly Agree” and “Agree” and conclude that on the pre-test, 22% of teachers agreed with the statement, while only 10% agreed on the post-test - that is a 12% decline. Alternatively, we could do the same for the response categories associated with disagreement and add up the “Strongly Disagree” and “Disagree” categories and conclude that there was a 55% to 82% progression, or a 27% increase. We could thus report either of these two values in a similar way as Figure 4 does. This numeric representation reduces the complexity of the finding and has its place in the impact reports that Seenaryo is preparing, as well as for other public representation. However, the exercise we just went through also illustrates one of its limits: the need to reduce complex information to communicate only *one* aspect of the data. Another limit, which is relevant when communicating with funders and other entities that place a high value on evidence, is that this kind of representation does not yield any statistical conclusions. Whether or not a 27%

increase is statistically significant and/or what the generalized effect size of this change is are important pieces of information to evaluate the potential impact of an effect. That is, statistical significance indicates the probability that the reported difference between teachers' responses before and after the training could just have occurred by chance, or whether we can assume that the training indeed had an impact. Moreover, there are standard measures of effect size that facilitate comparisons across different items to establish what effect is relatively larger (e.g., Cohen's *d*). Thus, in order to demonstrate to entities interested in statistical evidence, we recommend conducting significance tests on the data. For ordinal data, the appropriate analysis would be an *ordinal regression analysis*. This analysis will tell us whether there is, overall, a difference in response patterns between pre-test and follow-up test. Running such an analysis on our running example shows that indeed there is [Chi-square (1) = 76.25, $p < .001$]. Thus, we would be able to report something like "After the training, a significantly lower percentage of teachers agreed with the statement "*I believe children learn best sitting behind a desk*". If such data analysis is not possible within Google sheets it could be outsourced, and possibly automatized, for a given data format.

Take-away: Statistical analyses can add information value to data presentation.

Conclusion

Ordinal data, the main outcome variable of Seenaryo surveys, require some specific considerations to make the most out of them. We have recommended some complementary visualization and data analysis approaches to aid internal and external data evaluation and interpretation.

Child Effects

Build organizational capacity to have child evaluations conducted to assess Playkit's effects

Seenaryo has so far evaluated changes in teachers' behavior, attitudes, and knowledge before and after using the Playkit. These were mostly *indirect* evaluations, meaning that teachers were asked about their experiences. Whether or not Seenaryo could also conduct *direct* teacher evaluations is a topic we cover in sections above. When it comes to evaluating the children, neither indirect nor direct evaluations have been a

part of Seenaryo's proximate goals, mainly because such evaluations seem challenging. **However, since long-term improvement in child core life skills are part of Seenaryo's North Star goals (Table 1), their evaluation is a key part to show the Playkit's effectiveness.** In the following, we review possible paths for evaluating children along these paths, and discuss.

Child evaluations within the Playkit project

While Seenaryo does not yet have a clear roadmap concerning *how* to evaluate children, they have a very clear idea of *what* to evaluate in children, which is an excellent starting point to tackle the *how*. As to the *what*, Seenaryo targets an improvement in children's core life skills, a set of seven skills encompassing cognitive and language, physical, and socio-emotional skills. Bolstering these skills is proposed as a pathway to increase the chances of long-term success in school and life.

The life skills targeted by Seenaryo are the following:

- Communication (*social & language*)
- Cooperation (*social*)
- Resilience (*emotional*)
- Self regulation (*emotional*)
- Fine & gross motor skills (*physical*)
- Decision making (*cognitive*)
- Problem solving (*cognitive*)

There are, broadly, three ways one can evaluate such skills.

- Option 1: Direct observation: Trained evaluators would observe child behavior in naturalistic or semi-naturalistic situations and rate their behavior and responses based on predefined criteria. Parts of such an evaluation can potentially be automatized (see automatic evaluation of speech quantity in Deliverable 2).
 - Advantages: This is arguably the closest reflection of actual behavior and a direct child assessment. This technique can be very fruitful if the to-be-observed behavior is quite specific and limited and can thus be realized without too much resources. For instance, a quantitative count of speech quantity is much more convincing than a teacher reporting that they think children talk more.
 - Disadvantages: This technique is limited to the assessment of overtly observable behavior as well as behavior that can be easily elicited within

natural interactions. Requires personnel to conduct the observations. Coding schemes for some of the life skills would need to be developed, and the coding - whether automatic or manual - would require heavy training and postprocessing

- All in all it is an impactful and direct, but resource-intensive option.
- Option 2: Test administration to children. Standardized test kits are used for trained personnel to deliver standardized tests to children.
 - Advantages: These are standardized and validated measures that can be accessed simply by buying a test kit. It is a direct assessment of children, and it can be used to assess overt behavior, but also other skills such as cognitive or communicative skills. Test administration and delivery are streamlined.
 - Disadvantages: Requires some (but not extensive) training and practice with the respective test kit, requires personnel to conduct and score the tests with each child.
 - All in all it is a reliable and direct, mid-resource option.
- Option 3: Caretaker surveys. Standardized surveys are used to ask teachers (or parents) about child skills
 - Advantages: These are standardized and validated measures that can be accessed simply by buying a test kit. Easy and quick, no additional personnel needed for assessment.
 - Disadvantages: This is not direct, but an indirect child assessment. Requires someone to score/evaluate the results.
 - All in all it is a reliable and indirect, low-resource option.

Based on this overview, we conclude that Option 1 is not the optimal solution, except for perhaps the evaluation of speech quantity and back-and-forths in speech productions between teacher and child. We therefore will mention some example solutions for Options 2 and 3, which could be used concurrently and depending on needs and availability.

Take-away: Standardized direct and indirect tests of child life skills are a low- to mid-resource option to reliably assess changes in life skills.

Examples of standardized evaluation measures

The examples chosen are all well validated and approved tools. We based our

selection on their match with the targeted life skills, but it is neither an exhaustive selection nor does our mentioning of one tool over an equivalent other tool necessarily indicate any judgment about their respective quality. We chose examples that fit part of the age ranges targeted by Seenaryo, but the same test rarely covers both kindergarten and elementary school age ranges. We also prioritized tests that have relevant translations. However, note that the tests do not necessarily test only those life skills targets (e.g., the proposed test on emotion contains more than just self regulation but not only) - depending on the test, you might be able to only test and score the aspects you are most interested in (which is not recommended in more clinical settings, but possible for your specific purpose). Also note that for those tests that do not have translations, you might need to inquire with the publisher whether you can translate and use the test for your purposes.

Test kits usually contain material for testing as well as instructions on test administration and scoring.

Example 1: The Ages and Stages Questionnaire

[This questionnaire](#) (ASQ-3, Squires & Bricker, 2009) is an indirect comprehensive caretaker screening (Option 3). With *comprehensive* we mean that it covers a broad range of life skills. *Screening* means that it is not an in-depth tool, but can rather pinpoint areas of growth. It covers ages 1 month to 5 ½ years, which corresponds to the younger ages you are interested in. The areas screened are communication, gross motor, fine motor, problem solving, and personal-social skills. In addition this survey has a [companion tool](#) (ASQ:SE-2, Squires, Bricker, & Twombly, 2015) targeting social-emotional development.

The ASQ is filled in by parents and/or caretakers and scored by trained coders. It is available in multiple languages including English and Arabic. An alternative tool covering a somewhat broader age range (~8 years) is, for instance, Parents' Evaluation of Developmental Status (PEDs, Glascoe, 1998).

We recommend this or a similar survey tool as an entry point to child evaluations. Below, we introduce three more targeted tests, which fall into Option 2, direct child assessment with a structured test. Such tests could be used to go beyond broad-range screenings in the future, for instance for focused subsets of children and/or if strong evidence for specific life skills is needed.

Example 2: The Peabody Picture Vocabulary Test/ Arabic Picture Vocabulary Test

[This test](#) (PPVT, Dunn, 2019) is a standard tool for assessing early receptive vocabulary development between 2-6 years of age. The kit includes cards with pictures (in physical or digital form) on which the administrator can assess various aspects of receptive vocabulary. An optional add-on is a growth monitoring tool, which can be used to easily monitor changes over different test administration time points. A version in Arabic has been developed (APVT, Khammash, 1995), and is likely available upon request (at least research in English did not reveal the commercial availability of this test).

We recommend this or a similar standardized test for more in-depth, targeted testing of language skill development.

Example 3: Social Emotional Response and Information Scenarios

[This test](#) (SERAIS, Kim & Tubbs Dolan, 2019) is a recently developed tool that uses a scenario-based format, where children are asked to report what they would do and feel in a variety of different social situations. This test is intended to capture information about social, emotional, and cognitive skills among elementary school children in conflict-affected settings, and the measure has been tested in Lebanon with a sample of Syrian refugee children.

We recommend this or a similar standardized test for more in-depth, targeted testing of socio-emotional skill development.

Conclusion

Standardized test kits and surveys can capture the improvement in the seven life skills concurrent with Playkit use. Our examples illustrate the range of possible measures, and can guide the Seenaryo team towards a program that includes relevant child evaluations.

APP DATA

Optimize app data to show both reach and the effects of Playkit delivery

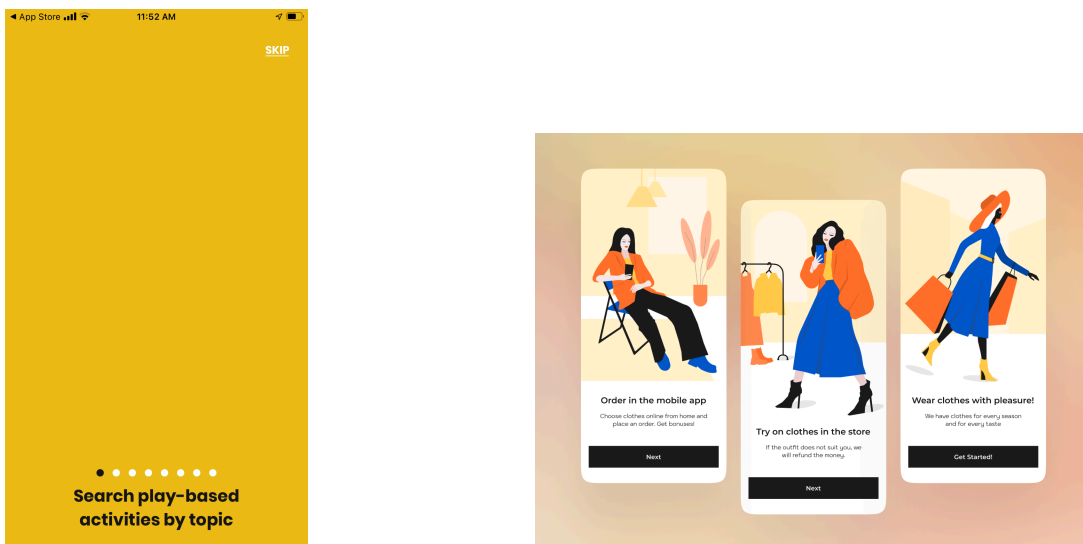
We see the Seenaryo app on the playstore as an important part of how you interact

with teachers and distribute content to them. As part of this recommendation, we reviewed the app as well as the data on its use to develop some recommendations and insights.

We started with reviewing the app from the perspective of a teacher/user and thinking about UI/UX and usability considerations. Following was the main feedback we had:

- Much of the functionality is on-point and consistent with what we would expect from such apps. The ability to personalize your language choices, search through content and save favorites are all implemented in the manner in which we would expect. Overall the app is easy to use and understand.
- Some of the visual experience could be enhanced. For example, there are several screens for onboarding/tutorials. But they have text bunched at the bottom and much empty space above. If you compare this to the on-boarding experience of most apps, you will find that they use some form of visuals in the empty space as part of engaging users. As a reference, we are comparing the Seenaryo onboarding screen with another reference below:

Figure 6



- The 'tips' section is particularly useful. But again it seems to be a sequence of written content. We would recommend either breaking it up with some visuals or

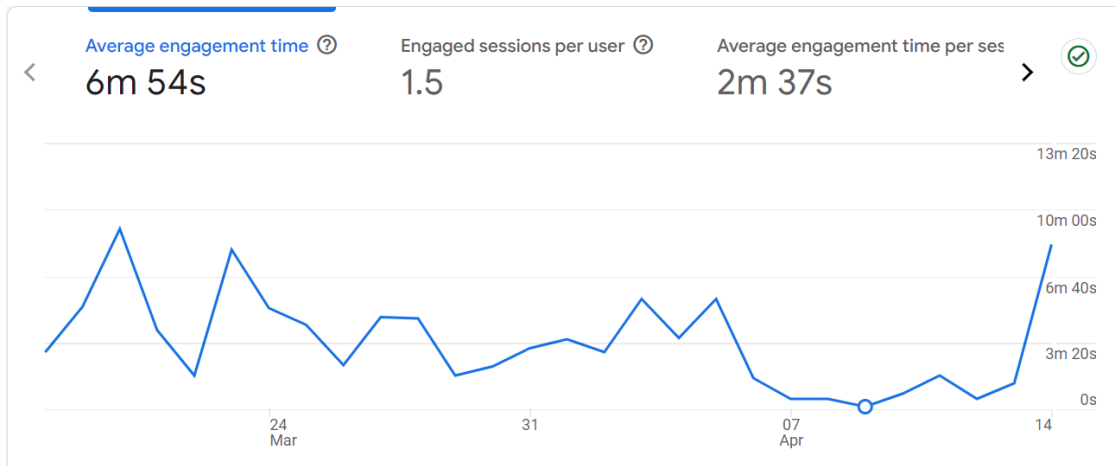
even using short videos since people are quite used to watching reels for basic instruction (though you should take into account internet availability/offline modes in such environments)

- We feel the biggest improvement area might lie in the sections on actual content. The labeling of lessons in terms of some basic parameters is quite helpful but some features that you should consider testing that might make the app content more useful:
 - Getting ratings and even advice from users who have tried out an activity. This should allow users to search popular activities or activities which have led to more engagement and discussion. This could also potentially increase the feel of community through the app with more active contributors getting some form of recognition.
 - We would also recommend additional labels of parameters that people might be interested in searching by. These could be labels like ‘sensory play’, ‘open-ended play’ etc. that might be particularly appealing in some contexts. This could also provide you with additional data on what people are searching for which would support your efforts in developing new content.
- Highlighting new/fresh content might also keep the app engaging and keep teachers thinking about the latest innovations in the space.
- Defining activities in such a way that they encompass multiple fields: using language learning activities to also talk about civics and citizenship and the environment might be interesting ways of promoting learning and keeping it relevant.
- Perhaps some basic content (again reels might be the best medium) speaking about the basic ideas behind constructivism and the supporting pedagogy could be quite useful in the context of the app as well.

Looking through the Google Analytics, we’d also like to share some observations:

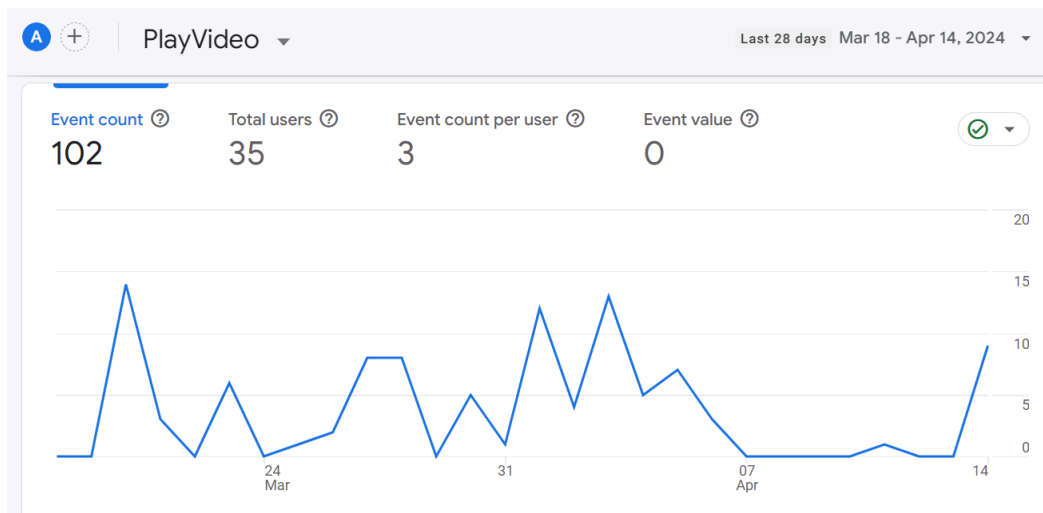
- The stability of the app is not a concern. We don’t see crashes on your iOS or Android app that would be reasons for the user experience to be affected. A small number of crashes on low-end Android devices should be expected.
- You have an average engagement time of 6 mins and 54s which seems a reasonable amount of average time to spend reviewing an app with such content.

Figure 7



- It would be particularly useful to look at which videos are being viewed. Here are the aggregate numbers:

Figure 8



And it is even more useful to look at breakdowns by video to judge which content is most popular. This should be used as your guiding reference on what content people are reviewing and what kind of additional content should be developed.

Figure 9

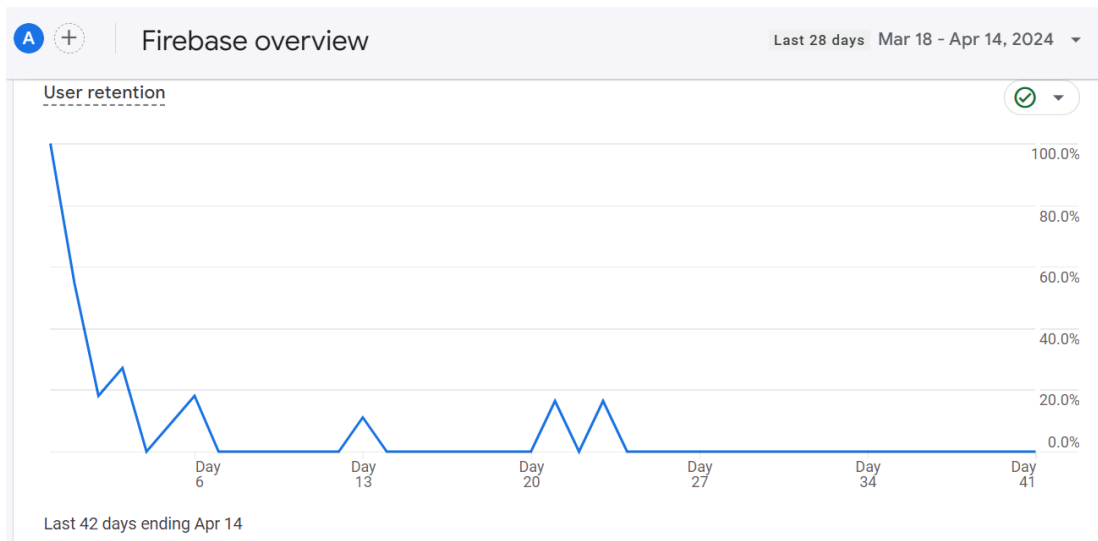
PlayVideo

Last 28 days Mar 18 - Apr 14, 2024

CUSTOM PARAMETER	EVENT COUNT	TOTAL USERS
(total) 45 items	102	35
Stepping Stones Grid	7	2
تغيير الأماكن	7	6
Balloon in the Box	5	5
أربعة جدران	5	5
البالون في الصندوق	4	2
Change Places	3	3
Horizontal Review	2	2

- There seems to be a pretty rapid drop-off in users beyond the first day or so. It seems like most new users see some initial value in the app but don't see a strong reason to keep coming back to it. We would recommend that some of the UI/UX feedback above, particularly around fresh content, would give people a reason to keep coming back and re-engaging with the app.

Figure 10



Overall, we would make the following recommendations around analytics:

- To make the firebase analytics richer, we would recommend measuring more events, particularly tracking exactly which activity is being tapped on and how

long it is being viewed to judge which content is the most useful.

- We would recommend using the analytics to benchmark engagement, average time and drop off rates and running ongoing experiments with some tweaks and feature additions to see which of these changes make improvements.
- We would recommend including Google Analytics and/or Mixpanel with the Firebase analytics for a richer set of analytics. We’ve found that their inclusion allows better tracking of user journeys and can help you judge where people are leaving the app.

NEXT STEPS

Moving along the continuum of progressively more advanced research presents many challenges, the first being high study costs coupled with funding gaps where grants don’t typically cover such costs.

More and more foundations (impact investors, in particular) are funding rigorous impact evaluations, but the journey to reach that level can be long and arduous, and organizations have to balance the need for evidence with the core activities of their program.

There is much Seenaryo can do in the short and medium term to strengthen and define and evaluate the impact of the Playkit - and those outcomes will support their long-term goals to positively buttress child development for those who learn through the Playkit. Such practices will be key to Seenaryo’s growth and ability to attract needed additional investments to scale and bring their innovation to more children in the region.

Below is our set of combined recommendations that we believe will strengthen Seenaryo’s path to scale. Once operationalized, these recommendations can serve as Seenaryo’s roadmap to improve the evidence base of their impact.

RECOMMENDATIONS (in no particular order)	TYPE	TIME
--	------	------

<ol style="list-style-type: none"> 1. Finalize and adopt your North Star Outcomes and accompanying North Star Metrics <ol style="list-style-type: none"> a. Based on the finalized and approved outcomes and metrics, update your Theory of Change and Update your Logic Model 	Impact Evaluation	1 month
<ol style="list-style-type: none"> 2. Draft a preliminary rough plan for an impact evaluation of North Star outcomes, socialize with partners, and fundraise to pay for an impact eval in the next year 	Impact Evaluation	6 months
<ol style="list-style-type: none"> 3. Consider adopting a more sophisticated survey tool to gain access to features that can make data collection and analysis more efficient and versatile. 	Data Strategy	3 months
<ol style="list-style-type: none"> 4. Hire a consultant for deep analysis of currently held data and produce data visualizations of impact (an Impact Report) 	Capacity Building	3 months
<ol style="list-style-type: none"> 5. Work with a partner to set up Qlik in such a way that leverages its features for the analysis of already collected data and provides a foundation for integrating and connecting new data sources to gain insights. 	Data Strategy	3 months
<ol style="list-style-type: none"> 6. Implement the the updated UCL surveys to get the most out of the current data-collection methods, which are indirect child assessments, but are a low-resource readily available option 	Operational Optimization	1 month
<ol style="list-style-type: none"> 7. Intentionally align the data strategy with Seenaryo's northstar goals and regularly monitor this alignment as goals evolve, and probe the effectiveness of data sources to provide meaningful insight on progress towards specific goals. To do this <ol style="list-style-type: none"> a. Create a 3-part organization of the data strategy and adopt it and implement it 	Data Strategy	1 month push and then Ongoing and Continuous
<ol style="list-style-type: none"> 8. Build agency familiarity with various child evaluation tools - to shift to tracking child outcomes <ol style="list-style-type: none"> a. ASQ-3 b. PPVT c. SERAIS d. IDELA e. Class Audio Collection f. Build behavioral effect tracking into the app where possible 	Capacity Building	6 months

<p>9. Leverage data from existing standardized tests, or develop a baseline workaround for how you might administer standardized tests to children</p>	<p>Impact Evaluation</p>	<p>1 year</p>
<p>10. Identify a university and/or other research partners and engage that partner to team up with you for an impact evaluation</p>	<p>Outreach & Advocacy</p>	<p>1 year</p>
<p>11. Follow, study, and learn from competitors:</p> <ul style="list-style-type: none"> a. IRC/SesameStreet/AhlanSimsim. Their findings of program effectiveness can support your building evidence of your intervention effectiveness. b. Lively Minds impact study. Watch for the release of the final report. Use their report (and referenced resources) and their model to take cues from how they constructed an evaluation study of their theory of change and how you can construct one for the Playkit 	<p>Competitor Analysis</p>	<p>Ongoing and Continuous</p>
<p>12. Start studying macro-level effects (and socioeconomic development indicators) using Playkit (Identify and track large-scale records of school performance, student performance in those schools, and teacher retention to look for evidence of temporal trends related to the adoption and increased use of the PlayKit, or to inform targeted outreach to schools and teachers.) to establish baseline metrics</p>	<p>Capacity Building</p>	<p>Ongoing and Continuous</p>
<p>13. Leverage behavioral data from the Playkit app in line with key questions and integrate feedback mechanisms that collect additional useful information for Seenaryo and potentially other teachers.</p>	<p>Data and App Strategy</p>	<p>6 months</p>
<p>14. Optimize effect-capture from app data by:</p> <ul style="list-style-type: none"> a. Adding persistent identifiers for teachers in the PlayKit app to track the behavior of individual users over time b. Connect app usage data to other data sources such as surveys and audio recordings. 	<p>App Data Strategy</p>	<p>6 months</p>
<p>15. Optimize App Performance for Max Engagement</p>	<p>App Data Strategy</p>	<p>6 months</p>

<ul style="list-style-type: none"> a. Enhance UX/UI interface to achieve higher user engagement b. Enhance "tips" with visuals or short videos as people are quite used to watching videos for basic instruction c. Build app capability for users to rate activity d. Build search tags that allow teachers to find "sensory play" options, etc. e. Work w/ developers to tag new content as position it and label it as NEW f. Measuring more events, particularly tracking exactly which activity is being tapped on and how long it is being viewed to judge which content is the most useful. 		
<p>16. Use the analytics to benchmark engagement, average time and drop off rates and running ongoing experiments with some tweaks and feature additions to see which of these changes make improvements.</p>	App Data Strategy	3 months
<p>17. Include Google Analytics and/or Mixpanel with the Firebase analytics for a richer set of analytics. We've found that their inclusion allows better tracking of user journeys and can help you judge where people are leaving the app.</p>	App Data Strategy	3 months

References

1. Glascoe, F. P. (1998). *Collaborating with parents: Using Parents' Evaluation of Developmental Status to detect and address developmental and behavioral problems*. Ellsworth & Vandermeer Press.
2. Khammash, S. B. (1995). *Omani Norms for the Arabic Picture Vocabulary Test (APVT) for Schoolchildren at the Elementary Levels*. Paper presented at the American Educational Research Association Annual Meeting, AERA/SIG International Studies San Francisco.
3. Kim, H. Y., & Tubbs Dolan, C. (2019). *SERAIS: Social Emotional Response and Information Scenarios: Evidence on construct validity, measurement invariance, and reliability in use with Syrian refugee children in Lebanon*. NYU Global TIES for Children, 10.
4. Squires, J., & Bricker, D. (2009). *Ages & Stages Questionnaires®, Third Edition (ASQ®-3): A Parent-Completed Child Monitoring System*. Baltimore: Paul H. Brookes Publishing Co., Inc.

5. Squires, J., Bricker, D., & Twombly, E. (2015). *Ages & Stages Questionnaires®: Social-Emotional, Second Edition (ASQ®:SE-2): A Parent-Completed Child Monitoring System for Social-Emotional Behaviors*. Baltimore: Paul H. Brookes Publishing Co., Inc.
6. Luo, L., Reichow, B., Snyder, P., Harrington, J., & Polignano, J. (2022). Systematic review and meta-analysis of classroom-wide social-emotional interventions for preschool children. *Topics in Early Childhood Special Education, 42*(1), 4-19.
7. Chaux, E., Barrera, M., Molano, A., Velásquez, A. M., Castellanos, M., Chaparro, M. P., & Bustamante, A. (2017). Classrooms in peace within violent contexts: Field evaluation of Aulas en Paz in Colombia. *Prevention science, 18*(7), 828-838.
8. Murphy, K. M., Yoshikawa, H., & Wuermli, A. J. (2018). Implementation research for early childhood development programming in humanitarian contexts. *Annals of the New York Academy of Sciences, 1419*(1), 90-101.
9. Ndiaye, L. G. (2022). Developing conflict resolution skills among pre-primary children: Views and practices of naturalized refugee parents and teachers in Tanzania. *Global Studies of Childhood, 12*(2), 159-169.
10. Murano, D., Sawyer, J. E., & Lipnevich, A. A. (2020). A Meta-Analytic Review of Preschool Social and Emotional Learning Interventions. *Review of Educational Research, 90*(2), 227-263
11. https://media.nesta.org.uk/documents/standards_of_evidence.pdf
12. <https://www.povertyactionlab.org/evaluation/effects-play-based-preschool-learning-program-rural-ghana>
13. Reference: Is My Ed Tech Tool Making a Difference? An Entrepreneur's Guide to Using Research to Improve Products and Measure Impact. New Schools Venture Fund - Tonika Cheek Clayton and Cameron White. Oakland, CA. 2018.
14. https://www.wested.org/area_of_work/learning-technology/edtech
15. Korat, O., Arafat, S. H., Aram, D., & Klein, P. (2013). Book reading mediation, SES, home literacy environment, and children's literacy: Evidence from Arabic-speaking families. *First Language, 33*(2), 132-154.
16. Shand, R., & Bowden, A. B. (2022). Empirical support for establishing common assumptions in cost research in education. *Journal of Research on Educational Effectiveness, 15*(1), 103-129.

17. Unicef. (2020). The Middle East and North Africa out-of-school children initiative Jordan country report on out-of-school children.
18. Mares, M. L., & Pan, Z. (2013). Effects of Sesame Street: A meta-analysis of children's learning in 15 countries. *Journal of Applied Developmental Psychology*, 34(3), 140-151