Pre-Analysis Plan

Investigating cognitive, non-cognitive and behavioral mediators for building literacy through edutainment in Kenya*

Jasmin Baier[1] Nicolas Bottan[2] Anushka Ghosh[3]

January 10th, 2025

Abstract

Educational television has shown promise for creating supplemental learning resources in the home and shaping attitudes in developing countries. We conducted a comprehensive evaluation of the effects of watching a new educational television program at home. Besides instructional content, a key innovation of the show is its objective to change children's mindsets about reading, gender attitudes and socio-emotional learning (SEL). Following a large-scale RCT, we have designed a lab study that investigates the learning mechanisms that are crucial in building literacy through educational television. In order to better understand if and how the show succeeds in building literacy skills, this study attempts to look closer at whether there is evidence that this program facilitates cross linguistic transfer of reading skills, and expands the RCTs exploration of potential mechanisms. It also evaluates the role of curiosity, distinguishing between trait curiosity and state curiosity, directed toward books, as well as the influence of executive functions such as inhibitory control, cognitive flexibility and working memory on literacy and SEL outcomes. Additionally, the study assesses the moderating effects of caregiver engagement on literacy and curiosity outcomes and explores whether knowledge retention from the show is essential for developing comprehension skills. By integrating qualitative and quantitative methods, the research seeks to identify active pathways linking the intervention to its observed effects.

[1] Baier: Oxford University; jasmin.baier@bsg.ox.ac.uk
[2] Bottan: Cornell University; nlb75@cornell.edu
[3] Ghosh: Busara Center for Behavioral Economics; anushka.ghosh@busara.global

* We thank Jonathan Karl, Gitanksh Sethi, and Lara Katumbi for drafting support. The project was approved by the Meru University Institutional Research and Ethics Review Committee (MIRERC) with approval number MIRERC/042/2024.

Timelines

August- September 2024	Preparation, pre-pilot, and pilot of Baseline survey instruments, sample selection, randomization, field outreach
October 2024	Baseline data collection
October 2024-November 2024	Lab experiments run
November 2024	Endline data collection
January- June 2025	Data analysis and research output preparation
Fall 2025	Dissemination of results and preparation of publication of research results

1 Introduction

Although access to primary education has increased drastically in recent decades, recent attention has focused on “the learning crisis” where, in 2019, 57% of children in low- and middle-income countries (86% in Sub-Saharan Africa) did not achieve minimal reading proficiency by age 10 (WorldBank et al., 2022). This has only been exacerbated by the Coronavirus pandemic (Tadesse and Muluye, 2020). Addressing these shortfalls in the short and medium term is difficult in light of the dearth of key resources, particularly quality teachers (UNESCO, 2022). In this context, educational technology (EdTech) could play an important role in bridging the educational gap. Interventions such as computer-aided learning have been shown to be highly effective at improving learning, but at a large cost - particularly in contexts that require investments in additional infrastructure (e.g., Muralidharan et al., 2019; Araya et al., 2019). Educational television is a promising alternative: It is low cost, can potentially reach millions, does not require guidance, and growing evidence suggests that it can improve children’s learning (Mares and Pan, 2013; Cherewick et al., 2021). For example, observational studies have found that TV shows such as Sesame Street improved learning in the United States and other low-income countries (Kearney and Levine, 2019; Mares and Pan, 2013). On the other hand, evidence from small randomized controlled trials of teacher-guided educational TV viewing at school has been found to increase learning in many low-income countries (i.e., Borzekowski, 2018). Baier et al. (2025) show that edutainment can indeed be beneficial for reading comprehension, which was the aim of the evaluated intervention. However, several of the hypothesized mechanisms did not appear to be impacted and reading comprehension displayed a potentially important differential effect by language spoken at home. This study therefore aims to build a deeper understanding of the evidence created by Baier et al (2025).

1.1 How this study builds upon previous field results

The evaluated intervention, an edutainment show titled Nuzo & Namia (N&N), aimed to develop language-agnostic skills known to facilitate reading comprehension. These include visualizing while reading, skimming, questioning, recalling, predicting, and making connections. However, language may significantly influence both engagement with the show and the encoding of transferred information. Specifically, language could impact children’s understanding of the processes required to cultivate these skills, their curiosity about reading fostered through information gaps, and their ability to practice beneficial intermediary behaviors. Additionally, it is possible that the skills were developed in a language-agnostic manner, but the language used in the assessment limited children’s ability to demonstrate them.

Given the RCT’s significant effects on fostering curiosity and the growing recognition of executive functioning (EF) as a critical factor in both socio-emotional skills and reading, this study also examines the show’s impact on three core EF components: inhibitory control, cognitive flexibility, and working memory. Furthermore, qualitative data collection underscored the pivotal role of parents in fostering children’s engagement, habit formation, and learning. This study therefore seeks to quantitatively assess the influence of parental involvement on these outcomes.

2 Background

Kenya uses an early exit education policy where the first three years of schooling is conducted in a national language, and a universal transition to English as the language of instruction is mandated in grade 4. The early-exit transitional model is where a first language (L1) is used for all subjects as the medium of instruction for three years before transitioning to English, a second language (L2) for most children, without explicitly promoting the long-term development of L1 (de Galbert, 2020). The goal of such policies is to develop early reading skills in L1 while teaching L2 (the language used as MOI for the rest of schooling) as a subject. Crawford (2004) notes that while no single policy fits all contexts, early-exit models are less effective overall than those supporting the L1 development for six years or more.

Reading comprehension can be seen as the product of decoding and listening comprehension (Hoover and Gough 1990) or more broadly as the combination of code-based skills and meaning-related skills (Lesaux and Marietta 2011). Code-based skills include phonological awareness, phonics, and oral fluency. Meaning-related skills include vocabulary and comprehension strategies. Cummins (1979) first proposed that oral language and literacy skills developed in the child's first language (L1) can transfer to the second language (L2) through a process he described as linguistic interdependence. This theory, suggesting that there is a common underlying proficiency across languages, has guided the research and practice in multilingual classrooms.

Based on the work of de Galbert (2017), it is possible that the Nuzo and Namia show can facilitate cross linguistic transfer. However, its effectiveness and required dosage are yet to be determined. Additionally, caregivers could play an important role mediating these effects and thus should play a central role in the theory of change.

The "simple view of reading" is often introduced to educators in professional development programs focused on the science of reading. This model effectively underscores the critical need for both decoding skills and linguistic comprehension in reading. However, given that literacy levels still remain critically low, especially in LMIC contexts, there is a need to gain a deeper and more nuanced understanding of how children read better. Research over the past 35 years has deepened our understanding of reading, leading to several key advancements, leading to the Active View of Reading Model (Duke and Cartwright, 2021) that has the following tenets:
(1) Reading challenges can arise from a variety of factors, not all of which are directly related to decoding or listening comprehension, as originally suggested in the simple view;
(2) Decoding and listening comprehension (or, as they are more commonly referred to today, word recognition and language comprehension) do not operate in isolation, but rather intersect in significant ways; and;
(3) There are other vital factors in reading that the simple view does not address, such as active, self-regulatory processes, which play a major role in reading success.

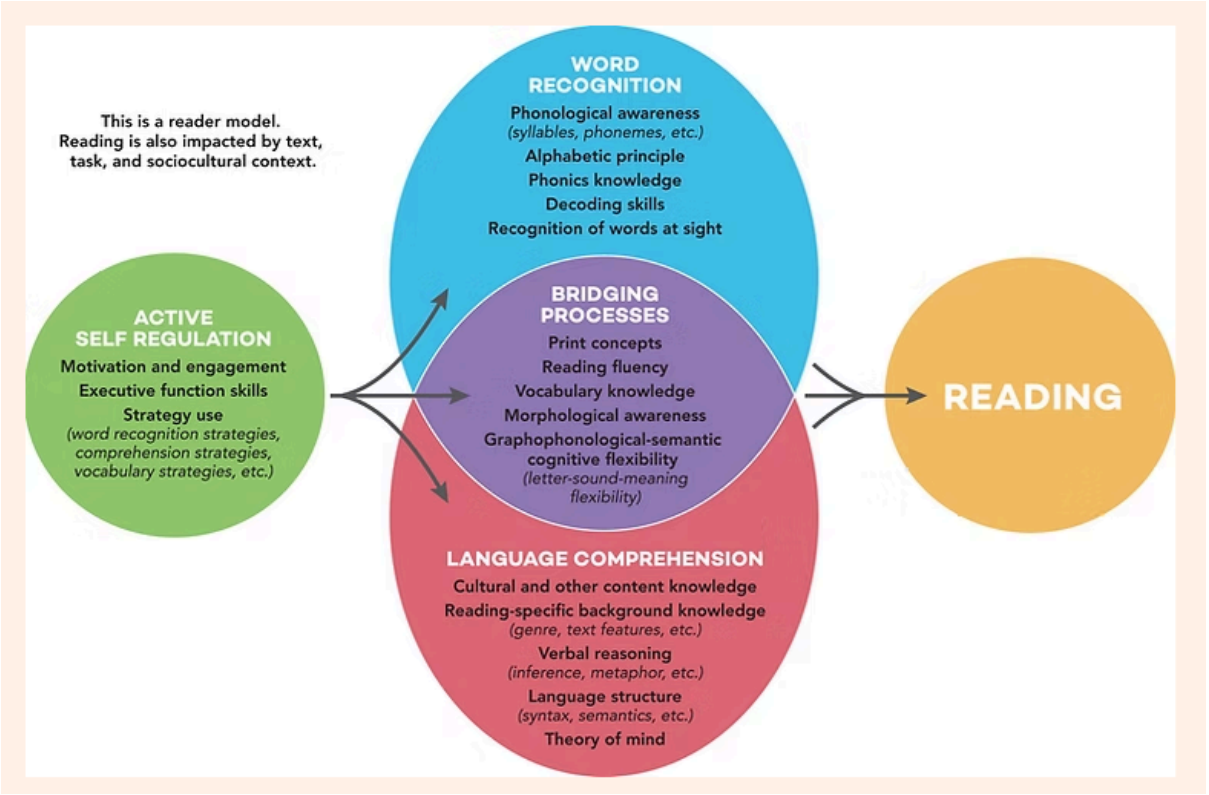

Figure 1: The Active View of Reading Model (Source: Duke and Cartwright, 2021)

To better understand the mechanisms that facilitate comprehension building, this study follows the Active View of Reading model (Duke and Cartwright, 2021) as demonstrated in table 1, to examine the relationship between active self regulation (divided into motivation and engagement, executive function skills, and strategy use) and both bridging processes and language comprehension; word recognition skills are not the aim of the evaluated show.

Moreover, we are interested in the mechanisms underlying the observed impact on curiosity from our at-scale RCT (Baier et al., 2025), as well as the quality of curiosity that is built. In Golman et al. (2021) and Loewenstein (1994), curiosity is characterized as "desire to fill 'information gaps'—specific unanswered questions that capture attention. Further, it has been found that people can be willing to incur costs to satisfy their curiosity (Eliaz and Schotter, 2010; Tasch & Houser, 2018). Given that the show seems to successfully increase a general measure of curiosity, yet we do not know whether this curiosity is also directed towards books, as the show intended, or if it is other elements of the show that affected this measure. Given the definition above it is possible that the show's adventures present children with a new and incomplete information set about other African countries and its inhabitants, which may be the source of children's curiosity. We are also interested in understanding whether the curiosity that was built may be related to children's willingness to incur costs towards indulging in an object of their curiosity, required that it does translate into a directed curiosity towards books.
Finally, we are interested in the underlying cognitive mechanisms that relate to children's curiosity. Neuroimaging studies have revealed that curiosity induction activates the anterior insula and anterior cingulate cortex (Jepma et al., 2012). Key functions of these areas include inhibitory control and cognitive flexibility. According to Cervera et al (2020), Inhibitory control, reward, and learning circuits

play key roles in driving curiosity. We will therefore test whether we find an indication for measures of inhibitory control and cognitive flexibility to mediate or moderate effects in our curiosity measure.

3 Research Objectives

This study aims to shed light on the following primary research questions:

1. Does Nuzo and Namia facilitate cross-linguistic transfer from the language in the broadcast to other languages, specifically from L1 to L2?
2. Might light caregiver engagement moderate effects on literacy skills?
3. Is the impact on curiosity a general (trait based) one, does it appear to be state curiosity and directed towards reading, or do we detect both?
4. Which mechanisms are salient in creating active pathways from input to impacting outcome measures?
    a. What is the role of three sub-components of Executive Function – inhibitory control, working memory, and cognitive flexibility – in mediating the paths from the intervention to curiosity and comprehension?
    b. Can induced state curiosity enhance knowledge retention about the informative aspects of the show, e.g. describe comprehension skills and book characteristics?
    c. Is knowledge retention of the show's information aspects (e.g. being able to describe comprehension skills and book characteristics) necessary for comprehension skills?

The secondary research questions explored here are:

1. What are the show's pedagogical practices in play during a 30-minute episode viewing experience?
2. Can we detect effects on reading comprehension skills already at a lower dosage of episodes, or only after completing the full season?

4 Research Design

4.1 Treatment Groups

Table 2 describes the experimental design and its relevance to Ubongo's TOC.

Table 2: Treatment Groups

Study Groups	**Iterations/ Description**	**Relevance to Programming/ Analysis**
Control Group: Placebo Show	This group will watch a placebo show content and not be exposed to the Nuzo and Namia treatment. At this moment, we propose showing the children a wildlife documentary to ensure no similarities in programming in literacy, SEL, or gender attitudes.	This group represents a control group against which we can compare results from the treatment groups to extract the impact on the sub-skills developed by the show.
Treatment 1 (T1): Receives the N&N show in Kiswahili without caregiver intervention	No iterations. This group watches the Kiswahili version of the Nuzo and Namia show.	This group helps us decode the cross-linguistic transfer component compared to the other treatment arms. It would be interesting to see if these children demonstrate skills transfer from Kiswahili to English or their mother tongue and whether we should expect that to happen to the original study sample.
Treatment 2 (T2)): Receives N&N in Kiswahili) with salient caregiver engagement nudges	This group watches in Kiswahili and receives some variations in the treatment, with caregiver engagement as the primary component. Narration, a recap at the end of the episode, and a short demonstration of building conversational literacy in the home while watching the show can drive caregiver engagement. Any other mechanism-related design will be incorporated during the research conceptualization phase.	We can decode the aspect of cross-linguistic transfer from this group by using the relative literacy scores between languages but controlling for absolute level differences between groups. We can also compare the results of having received caregiver intervention (absolute level differences).

4.2 Lab Experiment

4.2.1 Overview

The lab-in-the-field will essentially be a venue in a local community close to a public primary school from which we sample the households. Each group or treatment arm within the group will be seated in a room with a smart TV, where the show will be broadcast. There are six-episode viewings over six weeks.

4.2.2. Placebo[4]

We propose showing the kids in the Placebo control group a television show that they choose or a nature wildlife documentary that will not contain key programming around literacy, SEL, or gender. This placebo content also ensures that no harm is done while exposing children to certain content. It is not educational in the same sense or parameters as that of the Nuzo and Namia show. Some options for the placebo control are. It is important to note here that the show needs to be engaging enough to encourage attendance in the lab sessions but also not be harmful for the audience while trying to retain a true placebo. For this reason, we offer alternatives to a nature and wildlife documentary show from the list of preferred television shows that children in a previous randomized controlled trial with Ubongo specified. The top show from this list is Tom and Jerry which serves as a true placebo, does not contain harmful content, and also is considered engaging by the children.

4.2.3 Caregiver Intervention

The following aspects of caregiver intervention will be tested in the third treatment arm:

1. Engaging in meaningful conversations: Research shows that engaging in meaningful conversations with a caregiver leads to language and literacy development, especially in the earlier stages of life[5]. We aim to test whether this element is salient in the Nuzo and Namia show as well. We plan to include an element of a “recap” where the characters demonstrate doing a recall exercise with their parents on the show, at the end of each episode, and encourage children and caregivers watching the show to also talk to each other and recapitulate what the child did that day.
2. Teaching caregivers simple and easy-to-implement literacy development techniques: We would like to test whether equipping a caregiver with the skills to engage in simple literacy building activities with their child lead to better reading behaviors amongst children. An example of this would be to encourage them to read or tell a bedtime story with their children, demonstrate ways in which you can plan for richer conversations with your child on a busy day, and incorporate object-word association games with their children, amongst others.

4.2.4 Book and Print Recognition

Developmental literature suggests that reading habits are improved when a child is aware of the functions of a book and how to navigate reading experiences. In each group, we place different types of books in the lab room, including picture books, books with heavy texts, books of different languages, and some supplementary reading material aligned to the Nuzo and Namia show, to

[4] Please note that we will pilot each documentary and pick the one that is found most engaging by the children.

[5] Dupas, Pascaline, Seema Jayachandran and Mark Walsh. 2021. "Promoting Infant-Directed Speech in Northern Ghana." AEA RCT Registry. February 15. https://doi.org/10.1257/rct.7161-1.0

measure whether the children show curiosity towards different books and are able to describe the functions of the books.

4.2.5 Nuzo and Namia Episodes

The following episodes from Nuzo and Namia will be shown to the children in the treatment groups for the purpose of the study:

- NN104
- NN110
- NN101
- NN107
- NN106
- NN112

4.3 Sampling

The study involved a sample of 150 households selected from low-income communities in Nairobi, specifically from the Kibra, Kawangware, and Viwandani areas. Following authorization from the County Commissioner, the research team sought further approvals from the County Director of Education and the relevant Sub-County offices. Documentation of all study approvals is available for reference. Upon receiving the necessary approvals, the team engaged with selected schools within the sub-counties: Olympic and Toi Primary Schools in Kibra Sub-County; Kabiria Road and Kinyanjui Primary Schools in Kawangware, Dagoretti South Sub-County; and Jogoo Road and Makongeni Primary Schools in Makadara Sub-County. Using assistance from headteachers and class teachers, a team of trained enumerators visited these schools to obtain parent rosters, which were then used to contact and screen eligible households. Each household included one caregiver and one child as study participants. The households were subsequently randomized into one of three treatment groups, with randomization conducted at the household level.

4.4 Power Calculation

The experiment is designed to achieve 80% statistical power to detect a minimum effect size of 0.4, as measured by Cohen's d, which corresponds to a medium-to-large effect. If the treatment generates an effect of this magnitude, the study is sufficiently powered to detect the true effect with a high degree of probability. To further enhance our ability to detect effects, the primary analysis will utilize the pooled sample of the two treatment groups, comprising 100 households. In the pooled analysis, we achieve 80% power to detect a minimum effect size of 0.35. Notably, Baier et al. (2025) reported intent-to-treat effects (ITTs) ranging from 0.08 to 0.12 in a sample characterized by substantial two-sided non-compliance. Given this context, the chosen sample size is deemed appropriate for the controlled conditions of this study.

4.5 Data Collection

We plan to collect data through the following rounds:

1. Baseline: to establish baseline proficiency

2. Endline: At the end of 6 weeks to measure final impact
3. Exit Interview: A short exit interview is administered at the end of airing each bucket of skills to measure whether the dosage is sufficient.
4. Observational data: through 6 weeks for insight on pedagogy and engagement

5 Analysis

5.1 Outcomes and Mechanisms

I. Literacy (Family 1)

Primary Measures: We assess effects on literacy examining effects on three measures:

1. Reading comprehension with a focus on recall, visualization and inference.
2. Listening comprehension with a focus on recall, visualization and inference.
3. Visuospatial-linguistic skills (object word association and spatial recognition)
4. Summary score: These scores can be combined into a standardized summary literacy index (one for each language) employing Anderson (2008).

Reading comprehension and listening comprehension are the primary teaching objectives of Nuzo & Namia. All three measures are obtained employing subsections from the EGRA. Furthermore, each of these tests are conducted in three languages (English, Kiswahili, and Luyo) to test for cross linguistic transfer. Thus we will have a score for each language and measure (nine in total).

Secondary Measures: We compute EGRA listening and reading comprehension subscore identically constructed (percentage correct) as in the connected RCT studied in Baier et al. (2025) for comparability.

II. Socio-Emotional Learning (Family 2)

Primary Measures: For consistency, our primary SEL measures will be constructed in the same fashion as in Baier et al. (2025), enhanced by an additional curiosity and grit measure.

1. (Trait) curiosity score: As in Baier et al. (2025), items are added such that a higher score indicates higher curiosity. The score is then standardized to have mean zero and standard deviation of one (relative to control group).
2. (State) curiosity, directed towards reading: “Willingness to pay” for reading time measure: Number of clicks within allotted time of three (3) minutes
3. Confidence score: Baier et al. (2025), items are added such that a higher score indicates higher confidence. The score is then standardized to have mean zero and standard deviation of one (relative to control group).

Secondary Measures:

1. Grit: Time until the child stopped “trying”, i.e. stopped clicking the button in the effort task. The exercise allowed children to click for a maximum of three (3) minutes and documented the time if the child stopped before three minutes elapsed.
2. Combined SEL score: We follow Anderson (2008) to create an index which has a mean of zero and standard deviation of one, and is standardized relative to the control group. Given that the combined score mixes two constructs, we will furthermore conduct a principal component analysis as a more data-driven approach and as a robustness check.

III. Intermediate Behaviors (Family 3)

Primary measure:

1. Summary measure: As the primary measure in this family, we will construct a behavioral index following Anderson (2008)’s method for variance-weighted summary indices, without imputing missing index components.
    a. IB1: How often do you read outside of school, i.e. for enjoyment?
    b. IB2: How often do you look for or ask for books for enjoyment and not school work ?
    c. IB3: How often does your mum, dad or carer read books or stories to you?

Secondary measures:

1. Reading Behavior: Analysis on individual item IB1
2. Demand for books: Analysis on individual item IB2
3. Caregiver Behavior: Analysis on individual item IB3
4. Secondary Behavioral index: We will conduct polychoric principal component analysis (PCA) on the following likert-scale measures, after having standardized each item.

IV. Preferences (Family 4)

Self-reported preference between three distinct books; children are shown three book options in the lab room and asked to pick their *favorite* and *least favorite*. The data consists of 2 separate questions, which are both single select and have the same 3 choices (a text-heavy book, a ubongo book, a cartoon book).

Primary measures

1. Preference for reading: Indicator equals 1 if favorite book is either text-heavy book *or* ubongo book.
2. Light preference for reading: Indicator equals 1 if least favorite book is ubongo book *or* cartoon book.

Secondary measures

1. Strong preference for reading: Indicator equals 1 if favorite book is text-heavy book and least favorite is cartoon book.
2. Preference for ubongo: Indicator equals 1 if favorite book is ubongo book.

V. Information (Family 5)

Primary measures

1. Book and Print Recognition measure: Sum of correctly identified book elements and functions
2. Strategy Metacognition measure: Sum of correctly answered questions asking to describe two (2) skills taught in the intervention: recall and visualization.
    - Ability to describe and apply a comprehension skill separate from a text, based on the skill being demonstrated in episodes watched.

VI. Skill (Family 6)

Primary measures:

Executive Function

1. Inhibitory Control: Go/No Go Task
    - Domain_score_IC = Z(d-prime) - Z(RTM), d-prime and RTM are defined in secondary measures
2. Working Memory: 2-n back task
    - Domain_score_WM = Z(d-prime) - Z(RTM), d-prime and RTM are defined in secondary measures
3. Cognitive Flexibility & Creativity
    - Alternative uses task measure: Sum of eligible alternative uses the child could think of
4. Summary measure:
    - Unweighted, standardized composite score of three domains (Inhibitory Control (IC), Working Memory (WM), and Cognitive Flexibility (CF)), compute a subscore for each domain and then aggregate:

      Composite score =
      (Domain_score_IC + Domain_score_WM + Domain_score_CF) / 3

      Note that Domain_score_CF = Z(Alt_use_score)

Strategy use

5. Comprehension strategy application:
    - Both for visualization and recall, field officers responded with the following options:
        - Child was not able to visualize/recall at all
        - Child was able to visualize/recall 1-2 aspects/things [...]
        - Child was able to visualize/recall 3-4 aspects/things [...]
        - Child was able to visualize/recall 5-6 aspects/things [...]
    - For this measure, we will impute 0, 1.5, 3.5, and 5.5 respectively and take a simple mean across the two items (visualizing and recall).

Secondary measures:

Inhibitory Control: Go/No Go Task

1. d' (d-prime) = Z(Hit Rate) - Z(False Alarm Rate)
   Hit Rate = Correct targets / Total targets
   False Alarm Rate = False alarms / Total non-targets
2. Reaction time measure (RTM): Mean Reaction Time (RT) for correct Go trials plus mean RT for incorrect Go trials
3. Overall accuracy = (Correct Go + Correct No-Go) / Total trials
4. Mean Reaction Time (RT) for correct Go trials
5. RT variability (standard deviation)
6. RT coefficient of variation (SD/Mean)

Working Memory: 2-n back task

7. d' (d-prime) = Z(Hit Rate) - Z(False Alarm Rate)
   Hit Rate = Correct targets / Total targets
   False Alarm Rate = False alarms / Total non-targets
8. Reaction time measure (RTM): Mean Reaction Time (RT) for correct Go trials plus mean RT for incorrect Go trials
9. Overall accuracy = (Hits + Correct rejections) / Total trials
10. Mean Reaction Time (RT) for correct Go trials
11. RT variability (standard deviation)
12. RT coefficient of variation (SD/Mean)

Comprehension strategy application:

13. Minimum application measure: A binary variable which equals one if the child was able to visualize/recall at least 1-2 things (including children that were able to do more), and is 0 otherwise
14. Highly skilled application measure: A binary variable which equals one if the child was able to visualize/recall 5-6 things, and equals 0 otherwise

VII. Social interactions (Family 7)

Primary measures:

1. Social interaction index: We will construct this primary index following Anderson (2008)'s method for variance-weighted summary indices, without imputing missing index components, across the following items:
   a. "How often do you meet with friends to read (for enjoyment, not school books)?
   b. "How often do you ask friends for help if you have a problem to solve (problem related to studies or school)?
   c. "How often do you ask your parents for help if you have a problem to solve (related to studies or school)?
   d. "How often do you talk to your friends about your feelings?
   e. "How often do you talk to your parents about your feelings?

2. Parent conversational interaction index: Given that these items correspond to those used in Ganimian et al (2024), we will follow their outlined approach in constructing this index, using Anderson (2008)'s method.
    a. Talk to you while doing an activity w/ you around
    b. Describe things to you when walking
    c. Pointed, named object and asked you to repeat
    d. Read/ looked at a book with you
    e. Told you a story
3. General adult conversational interaction: Given that these items correspond to those used in Ganimian et al (2024), we will follow their outlined approach in constructing this index, using Anderson (2008)'s method.
    a. Sing to you
    b. Play with you
    c. Read/ looked at a book with you
    d. Told you a story
    e. Described things to you

Secondary measures:

1. Secondary social interaction index: We construct an index using polychoric principal component analysis (PCA) on the following items, measured on a 4-point likert scale. We standardize all items before conducting PCA.
2. Secondary parent conversational interaction index: We construct an index using polychoric principal component analysis (PCA) on the following items, measured on a 4-point likert scale. We standardize all items before conducting PCA.
3. Secondary general adult conversational interaction index: For robustness checks, we will compute a simple sum across all items.

5.2 Implementation Research

This study is meant to illuminate the "why" behind the results observed on literacy, SEL, and gender. We aim to do this by capturing detailed observational data that provides a measure for instructional practices and engagement during a 30-minute episode viewing.

Activity

The Stallings Snapshot Tool is used to conduct timed observations at random intervals during the 30-minute episode airing. The Stallings Snapshot Tool incorporates an element of mixed methods approach where data on specific activities is captured quantitatively and supplemented with qualitative observations. The tool, originally designed for classroom settings (World Bank, 2017), has been adapted to suit the context of lab in the field setting, ensuring a comprehensive understanding of children's engagement throughout the episode. At each snapshot, researchers capture 15-second snapshots of children's behaviors and interactions with the show. The tool, originally designed for classroom settings, has been adapted to effectively assess children's engagement, expressions, and behaviors during home viewing. By conducting multiple snapshots, researchers gather diverse data points, providing a holistic understanding of children's engagement throughout the episode.

Instructional Design

We adapt certain subitems from the Teach Tool which capture:

• The time the show spends on learning and the extent to which children are on task,

• The quality of teaching practices that help develop students' socioemotional and cognitive skills, and

• Other aspects of the learning environment such as the accessibility of the physical environment, including the 'classroom' set-up and materials available.

As part of the Time on Task component, three "snapshots" of 1–10 seconds are used to record both the character's actions and the number of children who are on task throughout the observation.

Child behavior and engagement

Following each snapshot, researchers use the Viewing Observation Tool to record their impressions of pedagogy and engagement during the specific segment observed. This qualitative tool provides researchers with valuable insights into how the children respond to the program and interact with its content (World Bank, 2017). The observations were conducted immediately after each snapshot, allowing researchers to capture real-time responses and experiences.

Caregiver Engagement and Behaviors

We collect data on how a caregiver tends to engage and interact with their child during an episode viewing, mapped to specific random segments in the episode. This data is meant to uncover how a caregiver behaves during an episode viewing and whether there are any observable differences in engagement across different treatment groups.

5.3 Estimation and Inference

Our main analysis will estimate our intervention's intention-to-treat (ITT), which should be equivalent to the Average Treatment Effect (ATE), unless participants fail to attend sessions. Because students were randomly assigned to treatment groups, we can identify the causal effect by estimating the following model:

$$y_i = \alpha_0 + \alpha_1 Treat_i + \beta_1 y_{i,t=0} + \varepsilon_i$$

where y_i is an outcome for the child *i* (e.g., the comprehension index), $Treat_i$ is an indicator that equals one if child *i* was treated. We include the baseline value of the outcome variable $y_{i,t=0}$, when available, to improve statistical precision. Our primary parameter of interest is α_1, that captures the average treatment effect of the treatment assignment on the corresponding outcome.

Our primary analysis will estimate treatment effects on the pooled treatment groups (T1 and T2) compared to the control, to maximize power. Following, we will also report treatment effects for T1 vs Control and T2 vs. Control separately.

We will initially run treatment regressions on all primary measures in the outcomes and mechanisms sections. Treatment effects on secondary measures will be presented in the Annex unless we have reason to believe that they are more informative than primary measures, e.g. due to implementation peculiarities.

In order to explore the likelihood of our hypothesized causal pathways we will furthermore present correlations between (i) our two measures of curiosity, (ii) our two measures of curiosity and comprehension, as well as (iii) all mechanism and outcome pairs.

Alongside original p-values, we estimate sharpened q-values that control the false discovery rate (FDR) across outcomes within each of our outcome families, separately for primary and secondary outcomes (Benjamini, Krieger, and Yekutieli, 2006).
In line with the subheaders in "5.1 Outcomes and Mechanisms", families of outcomes are defined as follows:

I. Literacy
II. Socio-Emotional Learning
III. Intermediate Behaviors
IV. Preferences
V. Information
VI. Skill
VII. Social interactions

In addition to treatment effects and raw correlations, we will perform an exploratory mediation analysis and modeling approach following Cartwright et al., (2020), which uses executive function and reading comprehension data from university students to explore direct and indirect pathways between these skills and other measured mechanisms.

5.2.1 Language transfer analysis

In Baier et al. (2025) we learned that the effect on reading comprehension may have been reduced due to the fact that we measured the impact of the english version on the show on a sample in which most children lived in households where English is not spoken commonly at home. While this was intentional, we would like the analysis of this study to now respond to three hypotheses:

1. Children who live in multilingual households where English is not the most common language of communication struggled to understand the content in English, which hindered their ability to learn comprehension skills.
    a. If this hypothesis were true, in our current sample with mostly children from households where English is not the predominant language, we would expect the effect of the show watched and tested in the language they are more familiar with (i.e. Swahili, in which the show was watched in this study) to be much larger than our average effect size in the field RCT.
2. If the show is effective in building language agnostic comprehension skills, we would observe a transfer of such skills across various languages that the child interacts with such as their mother tongue (Luo) and English, despite not watching the show in Luo or English; that is, we would observe similar treatment effects in all languages. If skill building is not language agnostic, we would expect to find a significant effect on swahili comprehension measures, but a much smaller or no effect on English and Luo comprehension measures.

3. Thirdly, differences in household characteristics may have mattered for the difference in skill-building – If this hypothesis were true, we would expect to find a smaller effect than in the field RCT or no impact on reading comprehension of our sample in swahili comprehension measures.

5.2.2 Compliance measures

We take attendance of both children and caregivers every week. We consider three different compliance measures:

1. **Full compliance:** Attendance of >90%, that is at least 5 out of 6 weeks.
2. **Partial compliance:** Attendance of >50%, that is at least 3 out of 6 weeks.
3. **Non-compliance:** Attendance of less than 3 out of 6 weeks.

We will use all of these compliance measures to calculate and report Treatment-on-the-treated (ToT) effects alongside our main ITT estimates.

5.2.3 Heterogeneity Analysis

To explore heterogeneity, we will interact indicators for variables listed below with the treatment term for the following dimensions of heterogeneity.

- Language spoken at home
- Gender of the child
- Barriers to reading,
    - Proxied by item which measures numbers of books in household
- Socioeconomic Status (SES),
    - Proxied by caregiver education
- Caregiver Engagement, Mindsets, and Behaviors.
- Baseline executive function
    - Primary measures for inhibitory control and working memory as described above
- Baseline literacy
- Character relatability

Subgroup analyses will be conducted with appropriate adjustments for multiple testing to ensure robust results. For categorical variables, subgroup means will be compared, while for continuous variables, we will explore interactions linearly or categorize them into quantiles where relevant. We may also run analyses for other sources of heterogeneity in a more exploratory manner.

5.2.4 Mechanisms and Further Analysis

I. Comprehension

We will initially treat all of our mechanisms as intermediary outcomes, and run treatment regressions on all measures defined in “5.1 Outcomes and Mechanisms”. Subsequently, building on Duke & Cartwright (2021), we will in particular explore the importance of three channels which form “active

self-regulation" and through which, according to the "Active View of Reading Model", reading is affected:

- Curiosity as proxy for motivation and engagement,
- The executive function summary measure, and
- The metacognition measure as proxy for strategy use

II. Curiosity

We are interested in the concept of curiosity as a mechanism as well as an outcome in its own right and would like to explore the relationship between curiosity measures, demographic characteristics, and outcomes in greater depth. To achieve this, we will leverage data from both the initial field RCT and the lab study, noting that the RCT data has not previously been analyzed for this purpose before the submission of this Pre-Analysis Plan.

1. Exploring the nature of the measured curiosity increase

To examine the quality of curiosity fostered by N&N, we will explore correlations between state and trait curiosity. This is important to get an initial idea of the type of information and skill gaps the show may have drawn children's attention towards and led to the heightened curiosity we measure (Golman et al., 2021; Loewenstein, 1994). We will also analyze text responses to the following prompts:

- *What are two things you want to learn more about in the next weeks?*
- *What are some things from the videos you watched each week that interest you very much?*

This analysis is intended to shed light on the types of information and skill gaps the show might highlight for children. Following best practices in qualitative research (Braun & Clarke, 2006), the text data will undergo both human coding (using thematic analysis) and computational approaches:

- Thematic Analysis: Using a bottom-up coding strategy to identify emergent themes, followed by a top-down analysis to detect predefined themes (e.g., mentions of books/reading, geography, interpersonal relationships, or activities portrayed in the show).
- Topic Modeling: Latent Dirichlet Allocation (LDA) will be used to extract dominant themes, complemented by clustering techniques (e.g., k-means or hierarchical clustering on word embeddings) for semi-automated analysis. These approaches are particularly suited for small datasets, although additional data quality checks will determine whether LDA is indeed possible (Blei et al., 2003).
- Large-Language Model (LLM) Analysis: An openly accessible LLM will be employed to identify emergent themes. While useful for initial exploratory analyses, LLMs are limited in quantifying theme prevalence, necessitating subsequent human verification.

2. Curiosity and memory encoding

We will investigate how state curiosity predicts memory encoding by regressing two recall measures — Book and Print Recognition and a measure of episode recall — on state curiosity and a standard set of demographic variables. We will only use the recall question, not the visualization question, that

we described in the measure “comprehension strategy application”. This builds on prior findings linking curiosity with enhanced memory through dopaminergic modulation (Gruber et al., 2014).

3. Socioeconomic-status (SES) and curiosity

Decker et al. (2024) find that lower-SES adolescents often prioritize exploitation over (risky) exploration, limiting their learning potential. Given the conceptual overlap between risk attitudes and curiosity — both reflecting a willingness to seek information beyond the status quo — we hypothesize a similar SES-curiosity relationship:

- Heterogeneous Treatment Effects: Analyze heterogeneous treatment effects on curiosity by SES.
- Conditional Correlations: Examine baseline conditional correlations between curiosity measures and SES, both in the RCT and the lab study data.

This aligns with evidence that SES influences cognitive and motivational processes (Mani et al., 2013) and could extend our understanding of curiosity as a socioeconomically patterned trait.

4. Inhibitory control and curiosity

Curiosity has been linked to inhibitory control, reward processing, and learning circuits (Cervera et al., 2020; Kidd & Hayden, 2015). While our lab study's experimental setup limits direct investigation of these mechanisms, we will explore the relative predictive importance of inhibitory control for:

- State and trait curiosity measures.
- Comprehension outcomes.

We will employ a modeling approach inspired by Cartwright et al. (2020) to analyze relationships in baseline observations. Based on our treatment regressions we will further examine correlations between gains in inhibitory control and both curiosity and comprehension respectively.

5.2.5 Robustness Checks

We will conduct robustness checks to confirm the stability of our results. These will include:

1. Replacement Analysis for Missing Data: Use multiple imputation methods or exclude early dropout participants to test robustness against potential biases from missing data.
2. Enumerator Effects: Control for potential enumerator effects to ensure specific facilitators’ guidance or oversight does not influence results.
3. Caregiver Interaction Effects: Control for caregiver engagement level as a moderator to see if literacy and SEL outcomes are robust when factoring in caregiver involvement.
4. Baseline Imbalance Adjustment: Identify any baseline variables that are substantially imbalanced across treatment groups and adjust for these variables in the primary analysis. This involves re-running key models with these variables included as covariates to assess whether results remain consistent despite baseline imbalances.
5. Excluding Baseline Controls: Re-run treatment regressions without baseline controls to verify that the estimated treatment effects hold without adjusting for initial covariate levels.

6. Alternative Variable Constructions: Test robustness by using slightly different constructions of key variables (e.g., alternative scales or categorizations) to ensure that results are not sensitive to specific variable definitions.
7. No Mean Imputation for Missing Data: Conduct analyses without mean imputation for missing baseline data to confirm that results are robust to different approaches for handling missing values, using listwise deletion or alternative imputation methods where appropriate.

5.2.6 Attrition

To address potential attrition bias, we will first assess whether baseline characteristics differ between respondents who completed follow-up and those who did not, as well as between treatment and control groups. This comparison will help establish if attrition is systematically related to treatment assignment or other baseline factors.

The primary analysis will present results without adjustments for attrition, if we find reason to do so in our checks. To check for differential attrition, we will conduct two analyses to evaluate the nature and impact of attrition on our findings:

1. Attrition Rate by Treatment Assignment: We will estimate a model with an indicator for attrition as the outcome to assess if attrition rates differ by treatment status.
2. Attrition and Baseline Covariates: We will regress an attrition indicator on a set of baseline covariates, e.g. respondent age, education, baseline literacy scores, household assets, and household size. Marginal effects will be reported, and we will test if each marginal effect significantly differs from zero.

If differences in attrition between any treatment groups exceed two percentage points or are statistically significant, we will implement two additional robustness analyses:

- Inverse Probability Weighting (IPW): Using the predicted probability of missing data from the attrition model, we will apply IPW to estimate the primary model with adjusted weights. This analysis will follow the hypothesis testing procedures outlined above.
- Lee Bounds for Attrition Bias: Following the trimming procedure in Lee (2009), we will construct bounds on key treatment parameters to account for potential bias due to differential attrition.

These approaches will ensure that our treatment effect estimates remain robust under varying assumptions about attrition.

References

Alan, S., & Mumcu, I. (2024). Nurturing Childhood Curiosity to Enhance Learning: Evidence from a Randomized Pedagogical Intervention. American Economic Review, 114(4), 1173–1210. https://doi.org/10.1257/aer.20230084

Baddeley, A. D., & Hitch, G. J. (1974). Working Memory. In G. A. Bower (Ed.), Recent Advances in Learning and Motivation (Vol. 8, pp. 47-89). New York: Academic Press. http://dx.doi.org/10.1016/s0079-7421(08)60452-1

Baier, J., Barasa, L., Bottan, N., Ghosh, A. and Millrine, M., (2025). Shaping Minds Through Living Room Screens: The Impact of an Educational TV Show on Kenyan Children's Learning and Attitudes. Available at SSRN: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=5022687

Blei, David M., Andrew Y. Ng, and Michael I. Jordan. (2003). Latent dirichlet allocation. J. Mach. Learn. Res. 3, 993–1022.

Braun, V., & Clarke, V. (2006). Using thematic analysis in psychology. Qualitative Research in Psychology, 3(2), 77–101. https://doi.org/10.1191/1478088706qp063oa

Cartwright, K. B., Lee, S. A., Taboada Barber, A., DeWyngaert, L. U., Lane, A. B., & Singleton, T. (2020). Contributions of Executive Function and Cognitive Intrinsic Motivation to University Students' Reading Comprehension. Reading Research Quarterly, 55(3), 345–369. https://doi.org/10.1002/rrq.273

Cervera, R. L., Wang, M. Z., & Hayden, B. Y. (2020). Systems neuroscience of curiosity. Current Opinion in Behavioral Sciences, 35, 48–55. https://doi.org/10.1016/j.cobeha.2020.06.011

Crawford, J. 2004. Educating English Learners: Language Diversity in the Classroom. 5th ed. Los Angeles, CA: Bilingual
Education Services

Cummins, J. 1979. "Linguistic Interdependence and the Educational Development of Bilingual Children." Review of Educational Research 49 (2): 222–251. doi:10.3102/00346543049002222.
de Galbert, P. G. (2020). Language transfer theory and its policy implications: exploring interdependence between Luganda, Runyankole-Rukiga, and English in Uganda. Journal of Multilingual and Multicultural Development, 44(1), 1–19. https://doi.org/10.1080/01434632.2020.1788038

Decker, A., Leonard, J., Romeo, R. R., Itiat, J., Hubbard, N., Bauer, C. C. C., Grotzinger, H., Giebler, M., Gabrieli, J., & Imhof, A. (2024). Exploration is Associated with Socioeconomic Disparities in Learning and Academic Achievement in Adolescence. PsyArXiv. https://doi.org/10.31234/osf.io/ucrme

Donders, F. C. (1969). On the speed of mental processes. Acta Psychologica, 30, 412–431. (Original work published 1868)

Duke, N. K., & Cartwright, K. B. (2021). The Science of Reading Progresses: Communicating Advances Beyond the Simple View of Reading. Reading Research Quarterly, 56(S1). https://doi.org/10.1002/rrq.411

Eliaz, K., Schotter, A. (2010). Paying for Confidence: An Experimental Study of the Demand for NonInstrumental Information. Games and Economic Behavior 70, 304–324.

Golman, R., Loewenstein, G. F., Molnar, A., & Saccardo, S. (2021). The Demand for, and Avoidance of, Information. SSRN Electronic Journal. https://doi.org/10.2139/ssrn.2149362

Guilford, J. P. (1967). The nature of human intelligence. New York: McGraw-Hill.

Gruber, M. J., Gelman, B. D., & Ranganath, C. (2014). States of Curiosity Modulate Hippocampus-Dependent Learning via the Dopaminergic Circuit. Neuron, 84(2), 486–496. https://doi.org/10.1016/j.neuron.2014.08.060

Hoover, W. A., and P. B. Gough. 1990. “The Simple View of Reading.” Reading and Writing 2 (2): 127–160.

Jepma M, Verdonschot RG, van Steenbergen H, Rombouts SARB and Nieuwenhuis S (2012) Neural mechanisms underlying the induction and relief of perceptual curiosity. Front. Behav. Neurosci. 6:5. doi: 10.3389/fnbeh.2012.00005

Kalanthroff, E., Linkovski, O., Henik, A., Wheaton, M. G., & Anholt, G. E. (2016). Inhibiting uncertainty: Priming inhibition promotes reduction of uncertainty. Neuropsychologia, 92, 142–146. https://doi.org/10.1016/j.neuropsychologia.2015.11.020

Kidd, C., & Hayden, B. Y. (2015). The Psychology and Neuroscience of Curiosity. Neuron, 88(3), 449–460. https://doi.org/10.1016/j.neuron.2015.09.010

Lesaux, N. K., and S. H. Marietta. 2011. Making Assessment Matter: Using Test Results to Differentiate Reading Instruction. New York, NY: Guilford Press.

Mani, A., Mullainathan, S., Shafir, E., & Zhao, J. (2013). Poverty Impedes Cognitive Function. Science, 341(6149), 976–980. https://doi.org/10.1126/science.1238041

McBreen, M., & Savage, R. (2020). The impact of motivational reading instruction on the reading achievement and motivation of students: A systematic review and meta-analysis. Educational Psychology Review. Advance online publication. https://doi.org/10.1007/s10648-020-09584-4

Lee, D. S. 2009. “Training, Wages, and Sample Selection: Estimating Sharp Bounds on Treatment

Effects.” Review of Economic Studies 76 (3): 1071–1102.

Loewenstein, G. (1994). The psychology of curiosity: A review and reinterpretation. Psychological Bulletin, 116(1), 75–98. https://doi.org/10.1037/0033-2909.116.1.75

Oshin Vartanian, Ingrid Smith, Timothy K. Lam, Kristen King, Quan Lam, Erin L. Beatty, The relationship between methods of scoring the alternate uses task and the neural correlates of divergent thinking: Evidence from voxel-based morphometry, NeuroImage, Volume 223, 2020, 117325, ISSN 1053-8119, https://doi.org/10.1016/j.neuroimage.2020.117325.

Pelegrina S, Lechuga MT, García-Madruga JA, Elosúa MR, Macizo P, Carreiras M, Fuentes LJ and Bajo MT (2015) Normative data on the n-back task for children and young adolescents. Front. Psychol. 6:1544. doi: 10.3389/fpsyg.2015.01544

Tasch, W., & Houser, D. (2018). Social Preferences and Social Curiosity. SSRN Electronic Journal. https://doi.org/10.2139/ssrn.3211794

Verbruggen, F. & Logan, G.D. (2008). Automatic and controlled response inhibition: Associative learning in the go/no-go and stop-signal paradigms. Journal of Experimental Psychology: General, 137(4), 649-672.