Learning Through Living Room Screens:

Experimental Evidence from Kenya*

Jasmin Baier[1], Laura Barasa[2], Nicolas Bottan[3], Anushka Ghosh[4], and Mark Millrine[4]

[1]Oxford University, [2]University of Nairobi, [3]Cornell University, [4]Busara Center for Behavioral Economics

This Draft: March 19, 2025.

Abstract

We study the potential of educational television – delivered at scale and viewed at home without guidance – to improve foundational learning and socio-emotional skills in a low-income environment. Using a randomized encouragement design with 4,300 Kenyan children, we evaluate Nuzo & Namia, a new literacy-focused television show broadcast in English. The intervention increases reading comprehension by 9% of a standard deviation and curiosity by 11.7%, with significantly larger effects among children in predominantly English-speaking households. These impacts are driven by the show's informational content rather than changes in time use or parental investment. However, the program inadvertently reinforces traditional gender norms, a backfiring effect likely driven by low character relatability. The intervention is highly cost-effective, generating 6.77 learning-adjusted years of schooling per $100. We provide the first experimental evidence that mass media can directly shape children's learning and mindsets outside the classroom, underscoring both the benefits and potential unintended consequences of edutainment.

JEL Classification: I21, I25, J16, C93.

Keywords: Educational TV, EdTech, home watching, literacy, socio-emotional learning, gender attitudes.

*Baier: jasmin.baier@bsg.ox.ac.uk; Barasa: lauranelima@uonbi.ac.ke; Bottan: nlb75@cornell.edu, MVR Hall 3220, Ithaca, NY 14853; Ghosh: anushka.ghosh@busara.global; Millrine: markmillrine2017@gmail.com. We thank Johannes Haushofer, for important feedback and guidance, as well as Lara Katumbi, Cuong Pham Vu, Salome Njambi, Leonard Waweru, and Vaani Chopra for their excellent assistance and support. We are thankful to Julian Cristia, Ricardo Perez-Truglia, Ignacio Sarmiento, participants of the CSAE Oxford research seminar and CASEL for valuable feedback. This study was funded by EdTech Hub. Ubongo contributed funding to the initial information sessions, but otherwise did not support any research costs. The study was approved by the Maseno University Ethics Review Committee with approval number MUSERC/01202/23. Preregistry: AEARCTR-0013106. This paper has been conditionally accepted via pre-results review at the Journal of Development Economics under the title: *"Bridging Educational Gaps Watching TV at Home: The Impact of an Educational TV Show on Kenyan Children"*

1 Introduction

Despite large gains in access to primary education, in 2019, 86% of children in Sub-Saharan Africa did not achieve minimum reading proficiency by age 10 (UNICEF, 2022).[1] Education systems with early-exit language policies, where instruction shifts from the mother tongue to an official language (typically English or French) partway through primary school, may face distinct challenges in improving literacy outcomes (UNESCO, 2025). Addressing these shortfalls in the short and medium term is challenging, as it would require comprehensive curriculum overhauls, extensive teacher training, coaching, and materials (e.g., Keaveney et al., 2021), and the success of scaling these interventions depends critically on the quality of these inputs (Kerwin and Thornton, 2021).

Educational technology (EdTech) could play an important role in bridging the educational gap and easing the language transition. Recent work has shown that interventions employing computer-assisted instruction are highly effective in improving learning (e.g., Muralidharan et al., 2019; Araya et al., 2024; Lai et al., 2015; Cardim et al., 2023). However, these types of programs can face substantial scaling and implementation challenges as well – particularly in contexts like Sub-Saharan African countries that require significant investments in additional infrastructure (i.e., Ogembo et al., 2012).

Educational television is a promising alternative: it is low-cost, easily scalable, and already reaches millions of children across developing countries.[2] Unlike many digital tools, it does not require the guidance of a tutor or parent, making it especially accessible in low-resource settings. Observational studies found that popular children's programs like Sesame Street improved learning in the United States and in low-income countries (Kearney and Levine, 2019; Mares and Pan, 2013). In addition, there is evidence from small-scale randomized controlled trials of watching educational TV programming at school, usually guided and supervised by teachers, that have been effective in improving learning (e.g., Borzekowski, 2018; Moran et al., 2023). However, there is limited causal evidence on the effectiveness of over-the-air educational television when viewed at home.

In this paper, we study whether educational television, delivered in a naturalistic setting (i.e., unguided, at-home viewing), can improve foundational skills and shift attitudes at scale. We conducted a pre-registered randomized encouragement design to study the causal effect of watching Nuzo & Namia, a novel educational television show broadcast over-the-air in Kenya.[3] To our knowledge, this is one of the first studies obtaining experimental estimates

[1] Minimum proficiency is defined as the ability to read a simple text *with comprehension*.

[2] For example, content produced by Ubongo regular reaches over 24 million children in Africa.

[3] The show was created by Ubongo between 2022 and 2023, the leading producer of educational television content in Africa. Ubongo is a non-profit providing a free over-the-air educational television channel with presence in 41 African countries, producing educational content in 12 languages.

of unguided, over-the-air educational television in a home setting. The main educational objective of the show is to improve literacy among six- to nine-year-olds, specifically focusing on comprehension skills. Reading comprehension is critical for breaking the cycle of learning poverty (UNICEF, 2022), yet it is the skill that Kenyan children find most challenging to acquire between Grades 1 and 3.[4]

Other innovative aspects of the show include its focus on shifting children's mindsets about reading, promoting various forms of socio-emotional learning, such as confidence and curiosity, and challenging traditional gender attitudes. These priorities align with the Kenyan school system's recent shift to a competency-based curriculum. This skills-oriented curriculum explicitly emphasizes socio-emotional skills such as creativity, self-efficacy (confidence in one's ability to perform challenging tasks), and "learning to learn" (curiosity and active pursuit of knowledge) (KICD, 2025).[5] Although the show is produced in many languages, in Kenya it was aired in English during the study – reflecting the country's early-exit language policy, in which the language of instruction shifts from a mother tongue to English in the fourth grade.[6]

One challenge in evaluating the causal effect of an over-the-air television show is the fact that it is accessible to everyone. We address this challenge employing a randomized encouragement design. To minimize the viewership of the new show among the control group, the show was aired by a network that in the previous year did not broadcast any Ubongo content, and the network did not advertise the new show at all for the purpose of this evaluation. We recruited 4,300 children from 346 public schools and sent SMS reminders to parents of children randomly assigned to the treatment group twice a week, noting the time and channel the show aired. Students in the control group did not receive any encouragement, information, or reminders about the show. The first season of the show aired from June 2023 and ran through March 2024. We collected endline measures for our outcomes soon after the season finished in April 2024 (about a year after our baseline survey).

The encouragement design was successful in significantly increasing take-up of the show.

[4]For instance, in Grade 1 44.29% of children met or exceeded expectations in comprehension—surpassing the 37.05% in reading aloud and 36.72% in word reading. By Grade 3 children's comprehension skills fall behind: only 43.17% meet expectations in comprehension, compared to 71.51% in reading aloud and 49.22% in word reading (Kenya Ministry of Education, 2020).

[5]The seven pillars of the competency-based curriculum encompass four core skills categories (Communication and Collaboration, Critical Thinking and Problem Solving, Creativity and Imagination, and Citizenship), and three primary mindsets to be taught across the new curriculum: Curiosity (Learning to Learn), Confidence (Self-efficacy), and Digital Literacy.

[6]English is traditionally taught as a subject in grades 1 to 3, with Kiswahili or the mother tongue spoken in the school's area as primary language of instruction. While Kenya's official languages are Kiswahili and English, about 61 indigenous languages are spoken throughout the country (Eberhard et al., 2025). There is some discretion at the school level about starting to use English as the primary language of instruction in the early grades.

Measuring compliance is difficult in this context because we cannot directly track viewing behavior at home. Therefore, we measured compliance in two main ways. First, we used traditional self-reports by children and parents. Second, to address potential demand effects, we developed novel questions based on character and episode recall. We find that children in the treatment group were 46 percentage points more likely to have reported ever watching the show, and 21 percentage points more likely to watching it consistently (i.e., at least 6 episodes). Recall-based measures are consistent with self-reports, although smaller in magnitude. For example, children in the treatment group are 35 percentage points more likely to correctly identify and name the main characters. We also present estimates of the local average treatment effect (LATE) of watching the show, using different measures of compliance to bound estimates.

Learning outcomes in our baseline and endline surveys are measured using the Early Grade Reading Assessment (EGRA). The ITT estimates suggest that the treatment increases children's comprehension score by 9% of a standard deviation. This effect is consistent with the show's explicit focus on comprehension skills and does not spill over to other literacy skills: we find near-zero effects on reading fluency (0.6% of a standard deviation), a skill not targeted by the program. The estimate compares favorably to more intensive interventions that have improved reading comprehension. For example, providing second graders in the US financial incentives to read increase scores by 18% of a standard deviation (Fryer Jr, 2011), or an intensive month-long school intervention proving teacher training and materials in Philippines increased comprehension by 15% of a standard deviation (Abeberese et al., 2014). Adjusting for potential two-sided non-compliance, the extensive margin estimates of the LATE for comprehension suggests an increase of 19.5-25.2% of a standard deviation for ever watching the show (for self-report and recall measures). However, the effects on comprehension are no longer statistically significant after conservatively adjusting for multiple hypothesis testing across all of our seven primary outcomes.

The treatment significantly improves children's socio-emotional outcomes, particularly by increasing their curiosity. Treated children exhibit a 11.7% of a standard deviation increase in their curiosity scores, whereas the effect on confidence is small (4.1% of a standard deviation) and statistically insignificant. This improvement in curiosity is economically meaningful and remains significant after adjusting for multiple hypothesis testing. It is comparable to the effects reported in Alan and Mumcu (2024), where a teacher training program raised curiosity by between 16.1% and 20.7% of a standard deviation.

Although the show aimed to challenge gender stereotypes and norms by portraying characters who defy traditional roles, the intervention unexpectedly reinforced more traditional gender roles among children. For example, the treatment significantly reduced children's

scores on gender role behaviors by 11% of a standard deviation. The point estimate is similar for gender bias (11.2%) and small but positive for gender stereotype knowledge (3.9%), though neither effect is statistically significant. This finding highlights how media interventions, while promoting learning, can inadvertently shape children's norms in unintended ways.

We examine heterogeneity across several pre-registered dimensions and use causal forests to identify the most relevant sources of variation in treatment effects (Wager and Athey, 2018). We do not find evidence of significant heterogeneity by gender, household size, SES, TV habits or baseline outcomes. For gender related outcomes, we find significant heterogeneity by caregiver's education, where the treatment worsens gender attitudes in low education households while it improves gender attitudes in higher educated households. There are also relevant differences for SEL and gender attitudes by the child's grade: SEL effects are strongest for third graders, while the decline in gender attitudes are concentrated among first graders.

One dimension of heterogeneity stands out across all three outcome domains and is also highlighted by the causal forests as particularly relevant: whether English is spoken at home. Recall that the show was broadcast in English, while the predominant language spoken at home in our sample is Kiswahili. English speaking children in the treatment group improve their comprehension score by 23% of a standard deviation, compared to 5.1% for non-English speakers at home (the difference is statistically significant). This effect is not driven by differences in compliance. Similarly, we find increases in curiosity of 27.8% of a standard deviation (compared to 7.1% for non-English speaking households). Although the effects on gender attitudes are negative in gender roles and gender bias scores for both English and non-English speakers at home, the estimates are significantly larger for English speakers. These findings suggest that the language chosen for airing the show is important for the overall efficacy of the show, and rather than improving preparedness for transitioning from a mother tongue to English, it may exacerbate existing achievement gaps.

We also explore mechanisms. First, we examine potential barriers for compliance and can rule out access to a functioning TV and prevalence of power outages as potential factors affecting compliance and outcomes. Second, we explore the potential channels through which our treatment could be affecting outcomes based on the framework proposed by La Ferrara (2016): preferences, time use, and information/skills. We find precise zero-point estimates for the effect of the encouragement on children's preferences (e.g., affect towards reading or characters, reading enjoyment, demanding books) or time use (e.g., total time spent watching TV). To proxy for the informational channel, we test children's ability to recognize the skills demonstrated in the show. Treated children are significantly more likely to identify the skills

being taught in different episodes. Third, we are also able to rule out alternative channels, like increases in caregiver investments or participation in their child's education. Children's ability to recognize skills, along with the finding that effects on literacy (comprehension) directly align with the show's learning objectives (i.e., teaching comprehension skills only), suggest that information is indeed the most likely mechanism driving our findings.

Finally, we present detailed cost-effectiveness calculations for a scaled-up version of our intervention that also makes use of encouragement, e.g. direct marketing. To our knowledge, this is one of the first papers to produce a detailed cost-benefit analysis of edutainment at-scale. We obtained detailed cost data from the show's producer. At 6.77 learning adjusted years of schooling (LAYS) per $100 based on gains in comprehension in a real-world setting, the intervention outperforms 91% of the 135 education programs analyzed in Angrist et al. (2025) in terms of cost-effectiveness. The cost per standard deviation is estimated at $18.48, or as low as $9.33 for English speaking households, which is significantly lower than $59.62 and $96.87 per standard deviation gain, for structured pedagogy and Teaching-at-the-Right Level (with technology) respectively – considered some of the most cost-effective foundational literacy interventions (Angrist et al., 2023). These estimates are conservative. They do not consider potential longer-term benefits from gains in comprehension, potential spillovers to other subjects, and they account only for one TV channel in Kenya, rather than all five countries where the show aired.[7]

Our paper contributes to a growing literature on scalable education interventions in developing countries (Glewwe and Muralidharan, 2016; Muralidharan and Niehaus, 2017; Banerjee et al., 2017; Angrist et al., 2025). In contrast to school-based teacher training interventions or computer-assisted-learning programs, our intervention requires no trained personnel, additional infrastructure, or direct supervision, and can be delivered at a very low marginal cost. Nonetheless, it generates comparable improvements in comprehension and curiosity. This suggests that home-based educational entertainment, while lower in intensity, can be valuable in contexts where delivery capacity is constrained. More broadly, our findings suggest that shows proven effective in small-scale, controlled settings can scale successfully and deliver meaningful benefits to the millions of children already engaging with educational television at home.

Our study extends the literature on socio-emotional learning (SEL) in children. Prior research has demonstrated that social and emotional skills are malleable (e.g., Alan and Ertac, 2018; Alan et al., 2019) and are critical for later-life outcomes, including academic achievement (e.g., Durlak et al., 2011; Taylor et al., 2017) and increased employment (Algan et al., 2022). In psychology, curiosity has been associated with enhanced life satisfaction and

[7]The show could easily expand to other countries in the region, further reducing the costs.

well-being (Schutte and Malouff, 2023; Lydon-Staley et al., 2020), as well as to improved job satisfaction and performance (Mussel, 2013; Kashdan et al., 2020). In a recent experimental study, Alan and Mumcu (2024) demonstrate that curiosity can be deliberately nurtured in young children. We build on this body of work by evaluating the effectiveness of a less intensive, media-based intervention for fostering curiosity at scale.

Our paper also contributes to a literature studying how media representations shape gender norms and behavior, that has typically focused on effects on adult women and young adults (Ferrara et al., 2012; Kearney and Levine, 2015; Jensen and Oster, 2009; Banerjee et al., 2019; Riley, 2024). Among children, previous research has found that similar interventions to ours shift norms in the intended direction (e.g., Kwan et al., 2020; Olsson and Martiny, 2018). Our study adds to this literature by studying how portraying gender-non-conforming role models affects different dimensions of children's gender attitudes, and can identify the influence of character relatability – a feature behind the psychological mechanisms of edutainment. Our findings suggest that using gender-non-conforming role models may backfire when characters are not relatable.

2 Research Design

2.1 Sample Selection

We conducted a school-level cluster randomized controlled trial. To ensure that our sample is as representative as possible, our sampling strategy consisted of 3 stages: (1) identified eligible counties and sub-counties, obtained a random draw;[8] (2) identified eligible schools within randomly selected sub-counties and obtained a random draw; (3) obtained final sample of children within eligible schools.

We used data on public primary school enrollment from the Kenya Basic Education Statistical Booklet 2019 (Kenya Ministry of Education, 2019) and data from the Kenyan Population Census to obtain sub-county school-level estimates of the number of eligible children. We defined an eligible child as one who was enrolled in grades 1-3 and had a TV at home, and an eligible sub-county as one that had at least 40 eligible children enrolled in a public primary school, and was not the only eligible sub-county in a county.[9] We then defined an eligible school as one that had an estimated minimum of 12 eligible girls and 12 eligible boys, so that there were at least 24 estimated eligible children. To reduce the chance of spillovers, we randomly sampled schools, such that each additional randomly sampled school was at least

[8] Kenya is divided into 47 counties and 314 sub-counties

[9] Nairobi was excluded due to a large likelihood of spillovers.

4 kilometers away from *any* other school already included in our sample.

With the help of the village chief, the village elders, and teachers, we obtained household contact information for the sampled schools. We then conducted a phone screening and recruitment exercise to find all households with children between 6 and 9 years and TV ownership. In Kenya, 81% of the population is reported to have access to TV (Murunga, 2021). The final sample had 346 schools (equally divided among treatment and control), with 2,198 children in the treatment group and 2,175 in the control group (12.64 children per school on average).

2.1.1 Randomization

Schools were randomly assigned to a pure control or treatment group where they received encouragement to watch the show (see details below). We employed a "multivariate stratified quadruple matching" technique for the randomization. While pair-wise matching has been the preferred approach to maximize balance and power in field experiments (Bruhn and McKenzie, 2009), recent work suggests that a more conservative approach employs the use of quadruplets (McKenzie, 2022). We followed this approach by employing a quadruplet clustered design.

Quadruplets were defined based on select baseline characteristics. Within each quadruplet, we randomly assigned two schools to the treatment group and two to the control group. The school-level characteristics used to define the quadruplets were: the share of female students in the school, asset score, grade (ranging from 1 to 3), caregiver education level, reading fluency score at baseline, cluster size, TV watching time, and household size. Quadruplets were matched based on clusters with the lowest Mahalanobis distance calculated using these characteristics.

2.2 Intervention

To study the causal effect of watching an educational TV show, we employed a randomized encouragement design. The show was transmitted free over-the-air TV; therefore, anyone with a TV could potentially watch the show. We took advantage of the fact that this was a brand new show. By design, there was no advertising for the show in Kenya. Additionally, the show was broadcast on KBC (Kenya Broadcasting Corporation), which had not aired any content produced by Ubongo (the show's producer) at least in the previous year. Therefore, the only information available about the show was provided through our initial parent information session and SMS encouragements. Only parents in treated schools received this information to encourage their children to watch the show at home. In this section, we pro-

vided details of the encouragement campaign and detail key elements about Nuzo & Namia, the TV show children were encouraged to watch.

2.2.1 Encouragement Campaign

We provided parents in the treatment group with various resources to encourage and remind them to having their children watch the TV show. Before the show started airing, treatment parents were invited to a group information session held in their local school or community center. Around 80% of parents attended these group sessions. The remaining parents that were unable to attend received a personal information session. One parent, the mother, typically attended. During the information session, parents were given an overview of the show and how the show is positioned to support their children's learning. To further increase buy-in among parents, a Ubongo fact sheet, testimonials from other parents who have engaged with past Ubongo shows, and an example of a scene from Ubongo was shown.

Throughout the show's run, parents were sent reminder SMS messages twice a week, 30 minutes before the show aired. After the show, parents also received an SMS question about the episode that had just aired, designed to elicit responses from the child. Responses to the SMS questions acted as an early warning sign to identify parents that were not following through with encouraging their children to watch.

2.2.2 The TV Show - Nuzo & Namia

The Nuzo & Namia show was developed by Ubongo and targets 6-9 year-old children with the primary objective of improving their reading comprehension. Additionally, the show targets socio-emotional learning, and gender attitudes. The plot revolves around 7-year old twins, Nuzo & Namia, who move into their grandmother's home with their family after the grandmother passes away to help the twins cope with their loss. In each episode, they take a book from a magic bookshelf and are transported on an exciting adventure to different African countries. In the episodes they learn about different places and cultures, and have to overcome different challenges. The series consists of thirteen 22-minute episodes, was produced in English, Kiswahili and Hausa, and was aired in five countries during its launch in 2023. In Kenya, the show aired twice a week on the Akili Kids channel and was broadcast in English. Each of the 13 episodes were broadcast multiple times throughout the period the show was aired.

The show incorporated several features to reach its learning objectives:

1. Reading Comprehension:

(a) Skill development and practice: The show presents learning strategies in an illustrative and engaging way that captures and holds children's attention (see the skill covered by each each episode Appendix B). The protagonists face challenges that require reading and comprehension to solve. In the process of solving these challenges, the protagonists model key strategies in early reading comprehension, such as predicting, questioning, clarifying and analyzing information.[10] During episodes, the protagonists occasionally break the fourth wall, asking the viewer questions and leaving time for responses. Thus, the show provides an opportunity for children to engage in informal learning along with the characters.

(b) Changing children's mindset: The goal was to change children's attitudes towards reading by creating characters who are relatable role models to children. Through character modeling behaviors such as reading books, they gain agency, confidence, and find motivation to improve their own reading skills.

2. Socio-Emotional Learning: The show focuses on two primary dimensions of socio-emotional learning: confidence and curiosity.

   (a) Confidence: The show fosters confidence in viewers through themes of risk-taking and self-belief. Characters step out of their comfort zones as they travel across African countries, embracing new cultures and challenges. Positive self-talk, such as "believe in yourself," reinforces self-confidence.[11] The series also highlights individuality and diversity through Namia's autism and the acceptance shown by others, emphasizing that confidence is essential, even when people are different.

   (b) Curiosity: The show fosters curiosity through the adventures of Nuzo and Namia, who eagerly ask questions and explore new environments. Their interactions—such as learning about different cultures and dances—highlight their desire for discovery. Bubelang, a magical guide, further nurtures curiosity by prompting imaginative thinking with open-ended questions. Blending factual learning with exploration, the show sparks a sense of wonder in children.

3. Gender attitudes: The show intends to shape gender attitudes by blending traditional and progressive roles. The mother emphasizes household organization, reflecting conventional norms, while the father challenges stereotypes by cooking and nurturing his children.[12]

[10] See https://www.youtube.com/watch?v=WiV638jzI84 for a short excerpt.

[11] See https://www.youtube.com/watch?v=yzCwToLRJb8 for an example.

[12] Another example is when the children discover old photos of their grandmother, showing her graduating from college and winning a swimming competition: https://www.youtube.com/watch?v=lCsofc7qhUg.

The show challenges gender stereotypes through its narrative, friendships, and roles. For instance, Namia often takes a leadership role, such as leading an interview in Ethiopia. Nuzo, on the other hand, is portrayed as nurturing, especially in supporting Namia after her autism diagnosis. These depictions break traditional gender norms, showcasing diverse character attributes.

2.3 Theory of Change

The first link in the theory of change is that the encouragement treatment increases the likelihood of watching the show. However, several factors may mediate this effect, potentially constraining viewership. These include limited access to a functioning television (although this barrier is eliminated through the eligibility criteria) or other constraints like unreliability of electricity supply.

Conditional on watching the show, outcomes can be affected through three main channels proposed by La Ferrara (2016): preferences, time use and information/skills.

- Preferences: The show targets children's preference for reading in many ways: modeling socially meaningful and relatable characters that promote positive attitudes (e.g., affect towards reading, gender roles, SEL). These changes in preferences could lead to other behavioral changes, such as children demanding more books. Changes in attitudes could directly affect some of our main outcomes, like SEL or gender attitudes. Preferences could affect literacy indirectly, through changing affect or interest in reading – thus increasing reading or other inputs to the educational production function (like books, or reading with parents).

- Time use: Watching the show could have also led to changes in time use, that could lead to changes in outcomes. For example, encouraging children to watch the show could have affected their general TV watching behavior. However, the direction of this effect is ambiguous ex-ante. To the extent that the encouragement crowded in or out the consumption of educational content it could increase or decrease learning. Increasing overall TV consumption could crowd out reading or studying time, thus decreasing learning.

- Information/Skills: The show conveys strategies for improving reading comprehension skills (e.g., visualizing, skimming, recalling, and predicting). These could have a direct effect on improving reading comprehension skills, as well as an indirect effect through decreasing the cost of reading, leading to more reading/practice overall.

Finally, there are some additional potential barriers to watching the show leading to learning gains. First, to the extent that practicing is an important mediator in improving literacy, the availability of books and reading materials at home might be a relevant barrier. Second, English fluency in the household could be important for take-up of the show, connecting with characters, and applying the skills taught by the show.

We also consider the possibility of other indirect effects induced by the encouragement treatment, such as peer influence and parental investments (Di Maro et al., 2024). For instance, the encouragement design could have signaled the importance of education or reading leading parents to increase education investments. Finally, the study deliberately shut down teacher-mediated channels by design. Recruiting and engaging study participants took place exclusively outside of schools, without involving teachers or advertising the show in school.

2.4 Data

2.4.1 Baseline and Balance

We conducted household surveys in May 2023 to collect our baseline sample of 4,373 children located in 346 schools.[13] Summary statistics are presented in Table 1. In our sample, 50% of the children are girls and is well balanced between grades 1 through 3. While all children have access to a TV, 91% of children have a TV at home, while the remaining 9% have access to a TV outside of their home (i.e., with a neighbor or family).

Table 2 shows that our randomization was successful. The treatment and control groups are well balanced and show small differences for the demographic, household characteristics, and baseline outcome variables. All differences in means between the treatment and control group are small economically. For instance, the treatment group is 2.5 percentage points more likely to predominantly speak English at home relative to the control group (p-value=0.226). The largest difference in baseline outcomes is 4.7% of a standard deviation for the SEL Index (p-value=0.191).

2.4.2 Endline and Attrition

We conducted our endline survey in April 2024, about a year after the baseline survey, and immediately after Nuzo & Namia finished airing in Kenya. Of the 4,373 children interviewed at baseline, we successfully followed up with 3,669 (83.9%). One potential threat to identification is the possibility that the treatment induced differential attrition. We explore this possibility in Table 3.

[13] These schools were located in six counties: Kajiado, Kakamega, Kiambu, Kilifi, Kisumu and Meru

Table 1: Baseline summary statistics table

	(1) Mean	(2) SD	(3) Min	(4) Max	(5) N	(6) Missing %
Demographics & HH Characteristics						
Child is female	0.50	0.50	0.00	1.00	4,373	0.00
Child in Grade 1	0.31	0.46	0.00	1.00	4,304	17.25
Child in Grade 2	0.34	0.47	0.00	1.00	4,304	17.25
Child in Grade 3	0.35	0.48	0.00	1.00	4,304	17.25
Hours watch TV per day	1.44	0.55	0.25	2.00	4,373	0.00
Watch 2+ hours per day	0.37	0.48	0.00	1.00	4,373	0.00
Watch Akili Kids Channel	0.44	0.50	0.00	1.00	4,373	0.00
HH size	5.59	1.86	1.00	10.00	4,373	0.00
HH has TV	0.91	0.28	0.00	1.00	4,373	0.00
HH has smartphone	0.58	0.49	0.00	1.00	4,373	0.00
HH Asset score	3.99	1.70	0.00	15.00	4,373	0.00
Caregiver has primary or less	0.63	0.48	0.00	1.00	4,373	0.00
HH speaks english	0.20	0.40	0.00	1.00	4,373	0.00
Nr books	13.98	13.48	5.00	50.00	4,373	0.00
Main Outcomes						
Literacy Index	0.01	0.98	-2.21	4.21	4,373	0.00
Comprehension Index	0.00	1.00	-1.68	1.68	4,373	0.00
Reading Index	0.01	1.00	-1.61	5.35	4,373	0.00
SEL Index	0.02	0.99	-2.89	1.33	4,373	0.00
Confidence score (z)	0.03	0.99	-2.48	1.15	4,373	0.00
Curiosity score (z)	0.03	0.99	-2.48	1.17	4,373	0.00
Gender Attitude Index	0.00	0.99	-3.22	3.91	4,373	0.00
Gender stereotype (z)	-0.02	0.99	-2.65	4.47	4,373	0.00
Gender roles (z)	0.01	1.00	-3.67	3.68	4,373	0.00
Gender bias (z)	0.00	0.99	-1.34	2.33	4,373	0.00

Notes: Baseline variable means and standard deviations in columns (1) and (2), minimum and maximum values in columns (3) and (4), and share of missing values in column (6).

We do not find evidence of differential attrition between treatment and control groups. The average attrition rates were similar in both groups (15.5% and 16.7%). The difference is economically small (1.4 percentage points) and is not statistically distinguishable from zero (p-value=0.266). This is robust to adding controls for school characteristics. Furthermore, we test whether baseline outcomes, and their interaction with the treatment indicator, are predictive of attrition. The joint test of significance suggests that differences in outcomes

Table 2: Baseline Balance

	(1) Treatment	(2) Control	(3) Difference	(4) p-value
Demographics & HH Characteristics				
Child is female	0.508	0.500	0.009	0.262
Child in Grade 1	0.302	0.314	-0.006	0.591
Child in Grade 2	0.353	0.331	0.022	0.149
Child in Grade 3	0.345	0.354	-0.016	0.147
Hours watch TV	1.430	1.441	-0.002	0.894
Watch 2+ hours	0.357	0.376	-0.016	0.279
Watch Akili Kids Channel	0.442	0.431	0.001	0.973
HH size	5.598	5.585	0.011	0.870
HH has TV	0.924	0.903	0.011	0.335
HH has smartphone	0.593	0.564	0.019	0.288
HH Asset score	4.020	3.960	0.010	0.790
Caregiver has primary or less	0.623	0.638	-0.006	0.655
HH speaks english	0.220	0.184	0.025	0.226
Nr books	14.044	13.914	0.019	0.976
Main Outcomes				
Literacy Index	0.018	0.000	-0.030	0.448
Comprehension Index	-0.002	0.000	-0.045	0.377
Reading Index	0.028	0.000	-0.012	0.700
SEL Index	0.039	0.000	0.047	0.191
Confidence score (z)	0.048	0.004	0.047	0.182
Curiosity score (z)	0.039	0.015	0.034	0.327
Gender Attitude Index	-0.009	0.000	-0.022	0.625
Gender stereotype (z)	-0.040	0.000	-0.030	0.395
Gender roles (z)	0.019	0.000	0.011	0.796
Gender bias (z)	-0.005	0.000	-0.045	0.445
Secondary Outcomes				
Reading comprehension (z)	-0.004	0.002	-0.040	0.325
Listening comprehension (z)	0.020	0.018	-0.036	0.535
Non-word reading (z)	0.037	0.001	0.003	0.917
Letter recognition (z)	0.042	0.015	-0.004	0.880
Oral fluency (z)	0.021	0.011	-0.023	0.559
Girl Gender roles (z)	0.016	0.000	0.005	0.891
Boy Gender roles (z)	0.008	0.000	0.010	0.727
Boy Gender bias: willingness to interact (z)	-0.239	-0.234	-0.031	0.441
Girl Gender bias: willingness to interact (z)	-0.324	-0.358	-0.003	0.947
Boy Gender bias: interactions (z)	-0.197	-0.207	-0.017	0.624
Girl Gender bias: interactions (z)	-0.276	-0.296	-0.017	0.651

Notes: Variable means for the treatment and control groups in columns (1) and (2). Columns (3) and (4) present the estimates for the adjusted differences in means and corresponding p-value for a significance test. The difference was calculated from an OLS regression of the treatment indicator on the balance variable, adjusting for quadruplet fixed effects. Standard errors are clustered at the school level.

and treatment status are not predictive of attrition (F-statistic=1.231, p-value=0.208). In Appendix Table A1, we further explore differences in baseline characteristics between attritors

Table 3: Differential Attrition

	(1) Treatment	(2) Control	(3) Adj. Difference	(4) p-value
Attrition	0.155	0.167	-0.014	0.266
	(0.362)	(0.373)	(0.012)	
Add school controls:			-0.013	0.290
			(0.012)	
Observations	2,198	2,175		
Joint test for selective attrition (F):			F=1.128	0.273

Notes: Mean attrition rate at endline (standard deviation below in parenthesis) for treatment and control group (columns 1 and 2). Column 3 presents the OLS regression adjusted differences, controlling for quadruplet fixed effects (school-clustered robust standard error in parenthesis). Column 4 presents the p-value for significance test of adjusted difference. Second row adds school-level controls. School-level controls (averages): share of girls, asset score, caregiver education, grade, reading score, HH size, TV watching time, and school size. Joint test for selective attrition presents the F-statistic and corresponding p-value of overall significance of an OLS regression where the outcome is an indicator for attrition and independent variables are all baseline outcomes plus their interaction with treatment, school-level controls plus their interaction with treatment and quadruplet fixed effects.

and non-attritors among treatment and control groups. On average, we do not find any statistically meaningful differences in baseline demographic characteristics or outcomes – suggesting that attrition is not significantly changing the composition of our treatment and control groups.

2.4.3 Main Outcomes

We have three distinct families of primary outcomes: literacy, socio-emotional learning (SEL), and gender attitudes. Following our pre-analysis plan, we study the effects in three steps to best account for multiple hypothesis testing. First, we construct three indices (Anderson, 2008; Schwab et al., 2020), one for each family of outcomes: literacy, SEL and gender attitudes. The indices have a mean of zero and standard deviation of one, standardized relative to the control group. We employed the same weights calculated for the indices at baseline when calculating endline indices to avoid changes in index composition from weighting.

Second, we study effects on our primary outcomes, as defined below. When studying these outcomes, we conservatively calculate q-values across all 7 primary outcomes to adjust for multiple comparisons following Benjamini et al. (2006). Third, we explore secondary out-

comes, and calculate q-values within each family of outcomes (e.g., within literacy outcomes and separately for gender outcomes).

1. Literacy: Our literacy module was designed following the Early Grade Reading Assessment Tool (USAID, 2016). The instrument was conducted *in English*, and included sections for letter recognition, non-word reading, oral fluency, reading comprehension and listening comprehension. Each individual component is studied as a secondary outcome.

   We construct two primary outcomes to measure literacy that capture distinct skills:

   - Comprehension index: combines percentage scores for reading and listening comprehension into a single measure.
   - Reading index: combines scores for: letter identification, non-word reading, and oral fluency.

   We benchmark our sample relative to a nationally representative sample of Grade 1 to 3 students in Kenya in 2021. Descriptive statistics are presented in Appendix Table A2. Overall, our sample performs quite similar or in some cases a slightly better than the nationally representative sample in 2021.

2. SEL: We use the RTI confidence and curiosity scale, which have been validated in both our East African context and in our age group (Jukes et al., 2021). A higher score reflects higher degrees of confidence and curiosity.

3. Gender Attitudes: One contribution of this study is the development and validation of measures to capture gender attitudes among children. These are coded such that a lower score indicates more rigid/traditional/biased attitudes or behaviors.

   For primary outcomes, we collected measures on three main dimensions: stereotypes, roles, and bias. Stereotype measures the level of stereotype knowledge that a child may have about a particular gender. Gender roles measure what the child actually does or is inclined to do (i.e., whether they engage in activities more or less aligned with their gender) and the character traits they exhibit. Gender bias measures the perceptions and behaviors of a child towards their own gender group and to those of their opposite gender. The overall measure of bias is obtained by the difference between the in- versus out-group for willingness to interact with the other gender and for their actual interactions. Details about the construction of these measures are available in Appendix C.

2.5 Identification Strategy

Our pre-registered analysis estimates the intention-to-treat (ITT) effects of our intervention on the treatment group. Because we employed a randomized school level clustered-treatment design, we can identify the causal effect by estimating the following model:

$$y_{is} = \alpha_0 + \alpha_1 Treat_s + \beta_1 y_{is,t=0} + \gamma_1 G_s + \gamma_2 \boldsymbol{X_s} + \epsilon_{is} \tag{1}$$

where y_{is} is an outcome for the child i in school s, $Treat_s$ is an indicator for school s being treated. We control for the baseline value of the outcome variable $y_{is,t=0}$, when available, to improve statistical precision. G is an indicator for the school quadruplet during randomization, and $\mathbf{X}$ is a vector of eight school-level controls that were used to generate the strata. Including these variables can increase the precision of or estimates.

Our coefficient of interest is α_1, that captures the intent to treat effect of receiving encouragements to watch Nuzo & Namia. Standard errors are clustered at the school level because this is both the level of randomization and stratification.

2.5.1 Defining Compliance

In our context, there is the potential for two-sided non-compliance: some students assigned to the treatment group fail to watch the show while some students assigned to the control group can watch the show. This is a direct consequence of the show being aired free over-the-air. We obtained various measures of take-up to estimate the local average treatment effect for treatment compliers (or LATE).

Measuring compliance in the context of watching over-the-air TV at home is difficult. In an ideal world we would like to track the exact number of episodes and engagement with each episode for each child. This would be completely unfeasible in a large-scale field experiment setting. Tracking viewing with audimeters (e.g., as Nielsen does with TV ratings) would not be technically feasible as they require cable boxes and the show is broadcast over-the-air.

We adopt a comprehensive strategy employing various measures for compliance. These measures are based on child self-reports and recall. We also obtain alternative measures that rely on parental responses to time use. Each measure has its own benefits and drawbacks. They capture different dimensions of compliance that we may care about. Thus, using different measures can provide useful bounds for the LATE estimates.

1. Child self-report: Prior work on the effect of watching a TV show outside the classroom mostly relied on simple self-reports from parents or children, asking whether they ever

watched a show or the frequency of watching the show (e.g., Mares and Pan, 2013; Moran et al., 2023). These measures are potentially subject to experimenter demand effects (De Quidt et al., 2018, 2019). If demand effects are larger among treatment children, estimates of the TOT would likely understate the true effects. We first replicate this approach to measuring compliance. Children were shown an image of the title screen of the show (containing the show's name and image of the two main characters) and were asked *"Have you ever watched this show?"*. If they responded "Yes", then they were asked: *"How many episodes do you think you watched?"*. The enumerator never mentions the show by name.

We construct two measures that likely overstate watching from these questions:

- Ever watched (self-report): an indicator that equals one if replies "Yes" to question *"Have you ever watched this show?"*. This measure captures intensive margin effects.
- Consistent watch (self-report): an indicator that equals one if reports having watched at least 6 episodes. This is because each skill is covered in two different episodes, for possibly watching each of the skills covered. This measure captures extensive margin effects.

2. Child recall: The previous measures have significant shortfalls. Besides potential demand effects, measurement error from imperfect memory of one's own activity is likely a problem. We constructed a novel compliance tool based on show recall that is asked to all children. Relying on recall also captures a different dimension of compliance: active engagement with the show (rather than simple viewership).

   First, children participated in a character recognition task by naming images of the three main characters. While Nuzo and Namia are easier to guess based on gender and context, the third character, Bubelang, has a unique name that is nearly impossible to infer. Responses were marked correct if they closely match the actual name.

   Second, children watch brief clips from 13 episodes. After each, they indicated if they recognized it and answered about the event immediately before or after the clip and what the characters learned. Enumerators were instructed: *For the following questions, show the child the episode that is going to be played on the screen. Please select "Yes" if the child clearly recognizes the episode and states the content similar to what is noted in the question. Please do not read the content/main takeaway to them.*

   One shortfall of this compliance tool is the possibility that children with higher ability (defined broadly) are likely to have better recall. This could potentially downward

bias our measure of true compliance. We tried to mitigate this as much as possible by design. For character recall, enumerators were instructed to mark responses as correct if they come close to the actual name. This task is objectively easier than episode recall. However, it is only an extensive margin measure of compliance. For episode recall, the excerpts were chosen in collaboration with the show's creators to select the most salient and important moment of each episode – thus, easier to recall.

We construct the following compliance measures based on recall. These can potentially be downward biased and may be correlated with child ability, thus overstating the TOT estimates:

- Character Recognition: an indicator that equals one if all three characters are correctly identified. Proxy for extensive margin effect.
- Consistent watch (identify): an indicator that equals one if the child correctly identifies the key moment for 6 or more episodes. Intensive margin effect.
- % Episodes identified: the percentage of episodes for which they correctly identify key moments.

3. Parent/Caregiver self-report: Parents/Caregivers are asked a diary-style question that can be used to construct a very rough proxy for watching the show. At baseline, around 75% of caregivers reported that their children watched TV on a given weekday, while 95% of children watched on the weekend.

3 Results

3.1 Effects of Encouragement on Compliance

We first evaluate the average effect of the randomized encouragement treatment on the compliance measures defined in Section 2.5.1. Table 4 presents the OLS estimates of equation 1, where odd-numbered columns only include quadruple fixed effects and even-numbered columns add school-level controls. A departure from equation 1 is that we do not control for the baseline outcome because these were only measured in the endline survey. Panel A presents estimates for self-reported measures, while Panel B present the recall-based measures as outcomes.

The encouragement treatment was successful in significantly increasing the viewership of Nuzo & Namia. The magnitude of estimates range from 149% to 353% in self-reported measures, and 223% - 324% in recall-based measures. Viewership in the control group was

Table 4: Effects of Encouragement on Compliance

Panel A: Self-Report:	(1)	(2)	(3)	(4)	(5)	(6)
	Ever watch		Consistent watch		% Episodes watch	
Treatment	0.462***	0.459***	0.213***	0.212***	0.203***	0.202***
	(0.019)	(0.019)	(0.015)	(0.015)	(0.011)	(0.010)
Mean Control	0.31	0.31	0.06	0.06	0.09	0.09
School Controls	No	Yes	No	Yes	No	Yes
R^2	0.284	0.290	0.136	0.136	0.207	0.209
Observations	3,669	3,669	3,669	3,669	3,669	3,669
Panel B: Recall:	Identify characters		Consistent watch (identify)		% Episodes identified	
Treatment	0.356***	0.354***	0.089***	0.086***	0.100***	0.098***
	(0.020)	(0.019)	(0.014)	(0.013)	(0.010)	(0.010)
Mean Control	0.11	0.11	0.04	0.04	0.05	0.05
School Controls	No	Yes	No	Yes	No	Yes
R^2	0.217	0.224	0.082	0.088	0.143	0.153
Observations	3,669	3,669	3,669	3,669	3,669	3,669

Notes: Each column presents estimates corresponding to an OLS regression of equation 1 for a different outcome or specification, though not controlling for the baseline dependent variable since they were only measured at endline. All estimates include quadruplet fixed effects. Even-numbered columns include school-level controls: share of boys, asset indicator, average grade, average reading score, school size. Cluster-robust standard errors at the school level in parenthesis. Significance: *** p<0.01, ** p<0.05, * p<0.1

fairly low considering that the show was available to everyone: 31.1% of children in the control group reported having ever watched the show. However, only 10.7% could accurately identify all three main characters. Children in the treatment group were 45.9 percentage points (p-value<0.001) more likely to ever have watched Nuzo & Namia than control children, as shown in Column (2) of Panel A in Table 4. Note that including school-level controls in even columns does not significantly change point estimates, though slightly improves precision. In columns (3) and (4) we present the effect for an intensive-margin measure of compliance: an indicator for watching at least six different episodes (self-reported in Panel A, and recalling episodes correctly in Panel B). The treatment significantly increased watching the show consistently by 21.2 percentage points (p-value<0.001).

Using self-reports can potentially overstate the true viewership and treatment effect of the encouragement. In Panel B of Table 4 we present estimates for different recall-based measures of compliance. When character identification is used, average compliance drops by 64% in the control group compared to the self-report for having ever watched the show.

However, the magnitude of the treatment effect only decreases by 22%. The encouragement increased character identification by 35.4 percentage points (p-value<0.001).

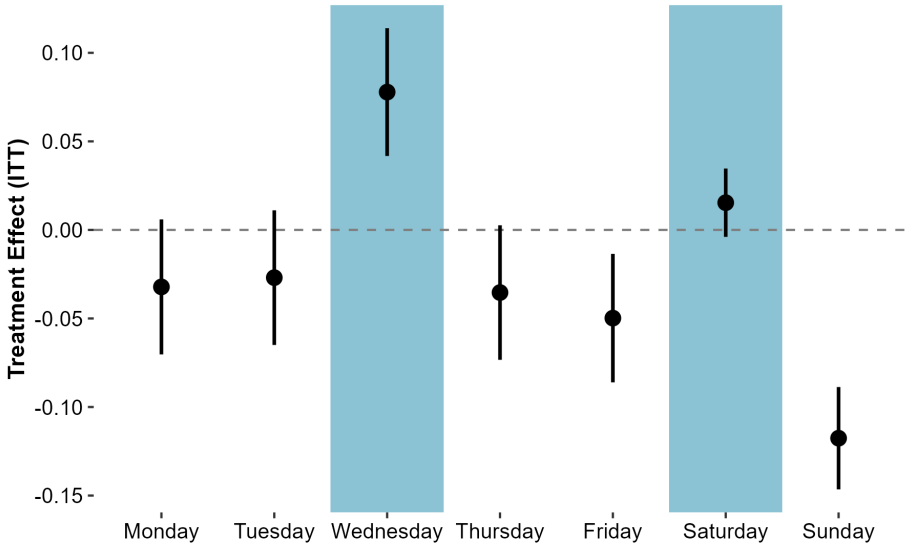
We replicate the intensity measure based on the number of episodes correctly *recalled* in columns (3) and (4) of Panel B. This is a significantly harder task, even for children who watched the show. Only 4% of the control children correctly identified six or more episodes, while treatment children were 8.6 percentage points more likely to have watched consistently according to this measure.

One concern with using recall-based measures is that they could be correlated with the child's ability. Indeed, we find suggestive evidence of heterogeneity in the effects of the encouragement by baseline literacy as presented in Appendix Table A3. While we do not estimate a significant difference for self-reported ever watching, all other measures indicate statistically larger effects for children with higher literacy at baseline. For example, the treatment effect on character identification for children with high baseline literacy was 6.8 percentage points larger than those with low baseline literacy (21.8% larger than for low baseline literacy children). The economic magnitude of this difference is larger for episode recall – a harder task, where the treatment effect was double in magnitude for high baseline literacy children.

We also find evidence that the treatment affected children's TV watching habits, as reported by parents. This behavioral measure is potentially less subject to either of the shortcomings of children's self-reported or recall-based measures. Figure 1 presents the coefficients and 90% confidence intervals for the treatment effects on the probability a child watched TV on a given day of the week. The show aired Wednesdays and Saturdays. The estimates suggest that children shifted their TV viewing days to Wednesdays and (slightly less) to Saturdays. In turn they reduced watching TV on Fridays and, very significantly, Sundays. Overall, we do not find that total TV watching significantly changed between the treatment and control children.

Overall, we find that the randomized encouragement was successful at increasing viewership of Nuzo & Namia among the treatment children. While precisely measuring intensity of take-up in this context is difficult, we find evidence that children not only ever watched the show, but did so enough as to recognize the key characters and the content of many episodes. Additionally, behavioral measures on TV watching behavior reported by parents are consistent with the large take-up of the show as reported by children. Compared to other message-based encouragement designs, ours was highly effective increasing compliance by 46-114% of a standard deviation. These effects are larger than studies messaging parents of preschoolers to promote literacy development which increased compliance by at most 34% of a standard deviation (York et al., 2019), or weekly reminders to primary school students in

Figure 1: Effect of Encouragement on TV Watching Days



Notes: Dots and lines correspond to point estimates and 90% confidence intervals of effect of encouragement on watching day. Days on which Nuzo & Namia aired are highlighted in teal. Estimates correspond to an OLS regression of an indicator for whether parent reported their child watches TV on a given day of the week (indicated along x-axis) as dependent variables on the treatment indicator, baseline dependent variable, quadruplet fixed effects and school-level controls. School-level controls: share of boys, asset indicator, average grade, average reading score, school size. Cluster-robust standard errors at the school level.

Perú to complete homework on an online platform that increased use at least once by 17% and 30% of a standard deviation using in-app notifications and SMS messages (Cristia et al., 2024; Aulagnon et al., 2025).

3.2 Effects on Main Outcomes

We next estimate the intent-to-treat (ITT) effects on our main outcomes. Following our pre-analysis plan, we present the OLS estimates of equation 1 in Table 5 for the three main indices that aggregate the primary outcomes for each family, and our primary outcomes. Odd-numbered columns present estimates only including the baseline outcome and quadruple (strata) indicators, while even-numbered columns add school-level controls. As pre-specified, to account for multiple comparisons we, on the one hand, aggregate measure into broad indices, and then for each primary outcome conservatively calculate q-values across the 7 primary outcomes (Benjamini et al., 2006).

Although we do not find significant gains in the overall literacy index, we find suggestive evidence of improvements in skills aligned with the show's learning objectives. Column (2) of Panel A indicates that the treatment increased the overall literacy index by 4.9% of

a standard deviation, though this estimate is not statistically significant (p-value=0.344). Recall that the show specifically targeted *comprehension* skills. Thus, we separately evaluate the effects on comprehension and reading in columns (3) to (6). Aligned with the show's learning objectives, we find that the treatment increased children's comprehension by 9% of a standard deviation (p-value=0.096), while the point estimate for reading is close to zero (0.6% of a standard deviation) and statistically insignificant. This suggests that increases in comprehension did not spill over to other measured skills.[14] The magnitude represents about one third of the baseline gap in comprehension scores by children with a caregiver having completed at least high school. However, the effect on comprehension is no longer significant after adjusting for multiple comparisons (q-value=0.225).

Consistent with the show's other learning objective for improving SEL, we find an increase in the overall SEL index driven by gains in curiosity. Panel B of Table 5 presents estimates for SEL. The treatment increased the SEL index by 9.1% of a standard deviation, on average (p-value=0.042, column(2)). Columns (3) to (6) of Table 5 shows the effects of the two main components of the SEL index: confidence and curiosity. Gains in the SEL index are driven by an increase in curiosity of 11.7% of a standard deviation and is highly statistically significant (p-value=0.009) even after adjusting for multiple comparisons (q-value=0.039). This effect is economically meaningful. For example, Alan and Mumcu (2024) studied a pedagogical training program for primary school teachers that was designed to nurture children's curiosity. They found that their treatment increased survey-based measures of curiosity by 16.1% and 20.7% of a standard deviation (for epistemic and scientific curiosity, respectively).

Estimates for gender attitudes suggest that the show might have backfired. Column (2) of Panel C in Table 5 suggests that the treatment lowered the overall gender index by 5.5% of a standard deviation, though imprecisely estimated (p-value=0.237). A negative coefficient reflects a shift towards more traditional (or biased) gender attitudes. The overall gender index captures different dimensions of beliefs and attitudes, which are explored in columns (3) to (8). We find a small but not statistically relevant increase in gender stereotype knowledge of 4% of a standard deviation.

However, we find strong negative effects on gender roles and bias (though bias is imprecisely estimated). On average, treated children reduce their gender roles score by 11% of a standard deviation. This estimate is highly statistically significant (p-value=0.011), and remains significant after multiple comparisons adjustments (q-value=0.039). Gender roles capture children's reported behaviors and traits. For example, a girl becomes more likely to

[14] This is further reflected when evaluating effects on the individual components of each index, as showing in Appendix Table A4. The coefficient for reading comprehension (0.084 SD) is similar in magnitude to listening comprehension (0.078 SD), though the latter is less precisely estimated. Point estimates for different reading skills (non-word reading, letter recognition and oral fluency) are all close to zero and insignificant.

Table 5: ITT Estimates - Primary Outcomes

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Panel A: Literacy			Standardized Scores					
Dep. Var.:	Literacy Index		Comprehension		Reading			
Treatment	0.043	0.049	0.080	0.090*	0.004	0.006		
	(0.055)	(0.052)	(0.057)	(0.054)	(0.050)	(0.046)		
q-value	-	-	0.375	0.225	0.942	0.898		
School Controls	No	Yes	No	Yes	No	Yes		
R^2	0.339	0.360	0.281	0.304	0.358	0.377		
Obs	3,669	3,669	3,669	3,669	3,669	3,669		
Panel B: SEL			Standardized Scores					
Dep. Var.:	SEL Index		Confidence		Curiosity			
Treatment	0.086*	0.091**	0.036	0.041	0.113**	0.117***		
	(0.046)	(0.044)	(0.042)	(0.041)	(0.046)	(0.044)		
q-value	-	-	0.460	0.419	0.048	0.039		
School Controls	No	Yes	No	Yes	No	Yes		
R^2	0.080	0.089	0.072	0.077	0.067	0.077		
Obs	3,669	3,669	3,669	3,669	3,669	3,669		
Panel C: Gender			Standardized Scores					
Dep. Var.:	Gender Index		Gender Stereotype		Gender Roles		Gender Bias	
Treatment	-0.052	-0.055	0.044	0.039	-0.108**	-0.110**	-0.097	-0.112
	(0.045)	(0.046)	(0.042)	(0.042)	(0.043)	(0.043)	(0.086)	(0.087)
q-value	-	-	0.411	0.419	0.048	0.039	0.411	0.346
School Controls	No	Yes	No	Yes	No	Yes	No	Yes
R^2	0.181	0.184	0.077	0.082	0.134	0.137	0.209	0.212
Obs	3,669	3,669	3,669	3,669	3,669	3,669	3,669	3,669

Notes: All dependent variables are standardized with mean zero, standard deviation of one relative to control group. Each column presents estimates for an OLS regression of equation 1 corresponding to a different outcome or specification. All estimates include controls for the baseline dependent variable and quadruplet fixed effects. Even-numbered columns add school-level controls (averages): share of girls, asset score, caregiver education, grade, reading score, HH size, TV watching time, and school size. Q-values correspond to False Discovery Rate (FDR)-adjusted significance levels across all seven mechanisms, following Benjamini et al. (2006). Cluster-robust standard errors at the school level in parenthesis. Significance: *** p<0.01, ** p<0.05, * p<0.1

report playing with dolls while boys become less likely to report playing with dolls. This backlash is not driven by a particular gender: Panel B of Appendix Table A4 suggests that the point estimates are similar for boy and girl gender roles.[15] At baseline, boys have signifi-

[15]Note that the variable for gender roles is not a simple average of the standardized scores boy and girl

cantly more traditional views on gender roles than girls, scoring 32% of a standard deviation lower than girls on the gender roles measure. Thus, the backfire effect on gender roles is about one third of the boy-girl gap.

Taken together, these results indicate that encouraging children to watch Nuzo & Namia on average increased children's curiosity, worsened gender beliefs, and are suggestive of an improvement in comprehension skills. It is important to highlight two points. First, these are intent-to-treat effects. As documented in the previous section there was imperfect compliance: some control children watched the show and not all treatment children watched the show. Additionally, effects could vary across sub-groups. For instance, because the show aired in English but a majority of children speak Kiswahili in the household (and at school for the first three years), watching the show could have differentially affected learning by language. We explore heterogeneity in the next section.

We present a detailed discussion of adjusting for imperfect compliance in Appendix Section D. Depending on the compliance measure used we estimate extensive or intensive margin effects (and potentially capture lower or upper bounds if self-report or recall-based). For intensive margin estimates, we find increases of 19.5%, 25.4% and -23.9% (25.2%, 33.0% and -31.1%) of a standard deviation for self-report (recall-based) measures of ever watching Nuzo & Namia on comprehension, curiosity and gender roles. As a measure of intensive margin effects we use an indicator for consistently watching the show (six episodes or more). We find estimates of 42.1%, 55.2% and -51.9% (104%, 137%, and -128%) of a standard deviation for self-report (recall-based) measures of consistently watching Nuzo & Namia. All estimates are significant at the 10% level, though only curiosity and gender roles are robust to multiple hypothesis testing adjustments.

The magnitude of the effects on reading comprehension compares favorably to those of substantially more intensive interventions like early grade reading programs. These programs typically include extensive teacher training, coaching, structured instructional materials, and regular student assessments. The median effect size on reading comprehension across eight evaluations of such programs in low- and middle-income countries was 43% of a standard deviation (see Graham and Kelly, 2019, for a review). We find that watching Nuzo & Namia, despite being only a fraction of the cost, produced comparable gains, ranging from 19.5 to 104% of a standard deviation (depending on the compliance measure used to adjust for non-compliance).[16]

gender roles, but rather they are averaged and then standardized. When averaging all scores, the variance for the control group becomes slightly smaller, explaining why the point estimate for the overall measure becomes slightly larger than the average for the sub-components.

[16]In Kenya, Tusome ("let's read" in Kiswahili) was a nationwide comprehensive early grade reading program costing over $88 million (around $5 per child).

3.3 Heterogeneous Effects

The encouragement treatment to watch Nuzo & Namia could have affected different sub-groups for several reasons. For instance, the show was aired in English and language could have been a barrier, either for take-up or for learning from watching the show. Alternatively, the caregiver's level of education could be important to improve the child's engagement or reinforce the teachings of the show.

To study heterogeneity, we follow the strategy detailed in our pre-analysis plan. We pre-registered 12 dimensions of heterogeneity measured at baseline.[17] We follow our pre-analysis plan by first estimating equation 1 by OLS for each sub-group. Second, we estimate causal forests (Wager and Athey, 2018) and obtain best linear projection (BLP) estimates (Semenova and Chernozhukov, 2021) of heterogeneous treatment effects.

We present OLS estimates for heterogeneous treatment effects in Appendix Table A5. In Figure 2 we present estimates for a sub-set of dimensions that were identified as important by the causal forest models (discussed below).[18] Overall, we do not find evidence of significant heterogeneity across most of the pre-registered dimensions. One dimension stands out as the only one having significant differences in treatment effects across all three families of outcomes: language spoken at home. This is particularly relevant in this context because the show was aired in English, though Kiswahili is the predominant language spoken in households. Around 80% of our sample reported speaking Kiswahili predominantly at home.

Columns (1) to (3) of Appendix Table A5 present estimates by language spoken at home. Children speaking English at home increase their comprehension score by 23% of a standard deviation (p-value=0.015) while non-English speaking children do not significantly improve their comprehension score (5.1% of a standard deviation). The difference in treatment effects is statistically relevant (p-value=0.042). Similarly, gains in curiosity are driven by English speaking children (27.8% vs 7.1% of a standard deviation), and the difference is also statistically significant (p-value=0.022). Both groups move towards more traditional gender roles, however the effect is larger for English speaking households (-20.3% vs -9.2% of a standard deviation) though this difference is not statistically significant. Additionally, we find that English speakers become significantly more gender biased (-32.1% of a standard deviation), however, this estimate is not robust to multiple hypothesis adjustments.

It seems that speaking English at home is a distinct dimension, and not capturing other attributes, like differences in socio-economic status, baseline levels of achievement, parental

[17] The dimensions are: gender of child, household size, speaking english in HH, child school attendance, caregiver education, SES, grade, whether watch TV with siblings or friends, whether watches TV at home, baseline level of outcome, number of books at home, urban vs non-urban schools.

[18] We selected the top three predictors of heterogeneity for each of the main indices as indicated in Appendix Table A9 excluding the strata (quadruplet) variable and baseline outcome.

education, wealth, or grade. We do not find significant heterogeneity by caregiver's education level, household assets, grade, or baseline outcome. One potential explanation for the gains among English speakers is that they disproportionately watched the show compared to non-English speakers. However, we do not find evidence of differential compliance (see Appendix Table A7). Moreover, we show in Table A8, that the English speaking subset of children remains balanced across observed characteristics. The heterogeneity by language suggests that the choice of language to air the show is critical for learning both language skills and SEL. At baseline, English speakers scored 28.5% of a standard deviation higher than non-English speakers. Thus, a estimates suggest that the show increased the attainment gap in comprehension between English and non-English speakers by approximately 75%, thus further exacerbating existing achievement gaps between English and non-English speakers outside of school (Hungi and Thuku, 2010).

We find that some other dimensions of heterogeneity matter, though not consistently across families of outcomes. For instance, columns (10) through (12) of Appendix Table A5 suggest that children in households with lower educated caregivers shift towards more traditional gender attitudes, whereas children with more educated caregivers adopt more *progressive* gender stereotype knowledge (a 14.9% of a standard deviation increase, robust to multiple testing adjustments). Whether the school is urban or not is another dimension suggesting that there are heterogeneous effects (columns (28) to (30)). Curiously, children attending urban schools increase their *reading* score (by 21% of a standard deviation) but not their comprehension score, while the increase in comprehension is larger for non-urban schools (though not significant after adjusting for multiple hypothesis testing). At the same time, children in non-urban schools experience a significantly larger increase in curiosity than children in urban schools. Finally, there is evidence of differences by school grade in SEL and gender attitudes (columns (34) to (39)). Increases in SEL are largest for third grade students, while shifts towards more traditional attitudes are mostly concentrated in first grade students.

Because we are exploring so many dimensions of heterogeneity across many outcomes, multiple hypothesis testing is a large concern in this context. Therefore, we use causal forests to flexibly explore heterogeneous treatment effects across all these dimensions of heterogeneity. This approach avoids the need to test many possible interactions or impose functional forms, mitigating concerns about multiple hypothesis testing and model misspecification. We estimate an honest school-clustered causal forest for each of the three main indices including quadruplets, baseline outcome and all dimensions of heterogeneity as potential explanatory variables.[19] Appendix Figure A1 shows that the estimated causal forests capture significant

[19]We do not include school-level controls because the model is able to highly predict treatment on them

heterogeneity for the three main indices. Appendix Table A9 shows the variable importance identified by the causal forests and Appendix Table A10 the estimates for the BLP.[20] Both present evidence consistent with our prior findings: they suggests that language at home seems to be an important dimension for heterogeneity, in particular for literacy and SEL.

4 Mechanisms

In this section we explore mechanisms following the framework outlined in our theory of change (see Section 2.3). We first examine potential barriers for take-up that can be correlated with compliance. Second, we explore the extent to which changes in time use, preferences and information/skills are driving our findings (La Ferrara, 2016). Finally, we explore the possibility that other factors than the show are driving our findings.

4.1 Barriers for Compliance

While we found that the encouragement design was successful at significantly increasing viewership of Nuzo & Namia, compliance could have been constrained by external factors. Our estimates suggest, that even with encouragement in the treatment group, take-up ranged between 46% using recall-based measures and 77% using self-reported measures (see Table 4, Column (2) in Panels A and B). Some of the observed non-compliance might be involuntary – for example, children could have had limited access to a functioning television, or other external constraints like electrical power outages which are frequent in Kenya.

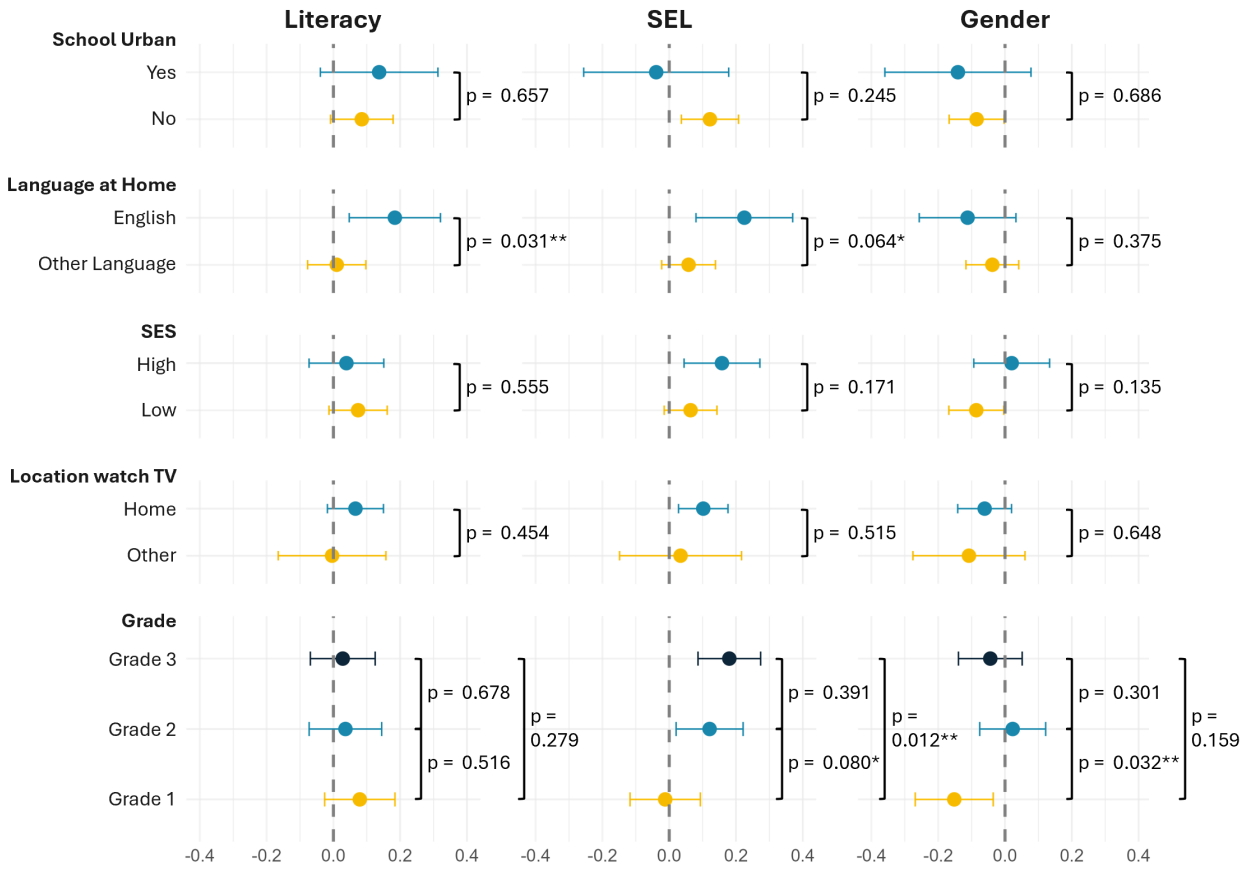
We do not find evidence that access to TV was a barrier to watching the show. First, we minimized this channel by design because our eligibility criteria required having access to a functioning TV. Indeed, around 87% of our sample had a working TV at home, while the remaining children had access to TV through a neighbor. One concern potential concern is that children without access at home could have had lower compliance. We test this directly in Appendix Table A11, where we study heterogeneity in compliance by whether the child has access to a TV at home. All the interaction terms are close to zero and statistically insignificant, suggesting that access to a TV was not a barrier to watching the show.

An alternative barrier to take-up is the reliability of the electricity supply. Although we did not collect survey data on outage frequency, we obtained monthly county-level data from Kenya Power. We computed the average System Average Interruption Duration Index (SAIDI), which captures the total duration of electricity outages experienced by the average

as they perfectly identify the treatment unit, thus we avoid overfitting by excluding them.

[20] As one would expect, quadruplet (i.e., the stratification variable) and the baseline outcome are the most important factors.

Figure 2: ITT Effects - Heterogeneity for a Selection of Dimensions Identified by Causal Forests



Notes: Each column corresponds to a different main outcome (index), each row presents different dimensions of heterogeneity. Dots indicate point estimates for ITT effect corresponding to a sub-group from an OLS regression that includes quadruple fixed effects, school controls and individual level controls. Lines are 90% confidence intervals. P-value for testing the equality of coefficients indicated for group pairs. Standard errors are clustered at the school level. Selection of dimensions drawn from top 3 attributes for each main outcome obtained from the variable importance indicated by the causal forests in Appendix Table A9 (excluding baseline outcome and quadruplets). See Appendix Table A5 for corresponding coefficients and standard errors, and all other dimensions tested.

customer, and divided the sample counties by the median value.[21] As shown in Appendix Table A12, higher exposure to outages is associated with smaller effects on extensive margin compliance – reducing the likelihood of ever watching the show and character identification by 11.4 and 13.5 percentage points (columns (1) and (3)), roughly a 20% decrease relative to low-outage counties. However, differences in intensive measures remain small and statistically insignificant, suggesting that frequent outages primarily deter marginal viewers, which might have occasionally watched Nuzo & Namia.

We do not find evidence suggesting that power outages significantly affected treatment effects on our primary outcomes. Appendix Table A13 explores heterogeneity by exposure to power outages. The interaction terms are economically small in most cases, and statistically insignificant. Although power outages could disrupt viewing, the repeated airing of episodes over several months likely allowed viewers multiple opportunities to watch any missed episodes.

Household characteristics, such as competition for TV access in larger families, and individual preferences, such as greater baseline enjoyment of reading, may also affect children's likelihood of complying with the encouragement. We conduct an exploratory analysis to identify the variables most strongly associated with compliance, replicating the heterogeneity analysis using compliance measures as outcomes. As shown in Appendix Tables A3, A15,[22] and A16, baseline literacy, reading affect, social interaction, and curiosity emerge as key predictors of compliance. Notably, baseline literacy consistently stands out as the strongest predictor across all three analytical approaches, particularly for intensive-margin measures, suggesting that the show was particularly appealing to children with some prior reading skills. This pattern is further corroborated by evidence that younger children (Grade 1) exhibit slightly lower compliance than older children (Grade 3) (see Appendix Table A3).

4.2 Preferences, Time Use & Information

We follow the framework proposed by La Ferrara (2016) to evaluate the mechanisms through which watching Nuzo & Namia could have affected our primary outcomes. As detailed in our pre-registration, we study the effect of the treatment on various proxy measures that capture changes in preferences, time use and information.

Watching Nuzo & Namia could have changed preferences and attitudes towards reading, as it is a central part of the show. We obtain different proxies to measure children's preferences and attitudes towards reading. First, we measure a child's stated preference for reading

[21] For reference, the average SAIDI for low versus high is 57 and 112, meaning that the average customer experienced 112 minutes of outage a month.

[22] Appendix Figure A2 shows corresponding targeting operator characteristics figures.

through an indicator for ever reading outside of school for enjoyment and ever asking for a book. Additionally, we construct an index capturing a child's positive affect toward reading, combining self-reported measures of reading enjoyment, interest, perceived ability, reading speed, and nervousness when encountering difficult books. Each component was assessed using a four-point Likert scale and administered twice, with the phrasing reversed in one instance, which allows us to test for response consistency.

We find little evidence that the intervention significantly changed children's preferences, as shown in columns (1) to (3) of Table 6. The treatment induced children to increase their reading for enjoyment by 2.5 percentage points. However, this effect is economically small and not robust to multiple hypothesis testing adjustments (q-value=0.598). Moreover, it may be subject to experimenter demand effects, making it unlikely to be a primary mechanism driving the observed outcomes. Similarly, the estimates for ever asking for books and for affect towards reading are both close to zero and statistically insignificant. These results suggest that changes in reading preferences and habits are not driving our main findings.[23]

Table 6: ITT Effects - Mechanisms

Mechanism:	(1)	(2) Preferences	(3)	(4)	(5) Time Use	(6)	(7) Information
Dep. Var.:	Ever Read for Enjoyment	Ever Ask for Books	Reading Affect Index	Watch 2+ hours TV a day	Parent-reported # days child watches TV a week	Mostly Watch Non-Educ. TV	Identifies any Comprehension Skill
Treatment	0.025*	-0.012	-0.016	0.033	-0.178	-0.047***	0.106***
	(0.014)	(0.014)	(0.030)	(0.025)	(0.124)	(0.016)	(0.026)
q-value	0.598	0.780	0.689	0.547	0.369	0.016	0.001
Control Mean	0.821	0.811	0.000	0.237	4.896	0.134	0.568
School Controls	Yes	Yes	Yes	Yes	Yes	Yes	Yes
R^2	0.012	0.024	0.070	0.091	0.168	0.069	0.153
Observations	3,669	3,669	3,669	3,669	3,669	3,669	3,669

Notes: Each column reports OLS regression estimates for a different outcome or specification. All regressions include quadruplet fixed effects and school-level controls. School-level controls (averages): share of girls, asset score, caregiver education, grade, reading score, HH size, TV watching time, and school size. The positive affect index is constructed following Anderson (2008) combining measures of positive affect toward: reading, combining self-reported measures of reading enjoyment, interest, perceived ability, reading speed, and nervousness when encountering difficult books. Identifies any comprehension skill is an indicator variable that equals one if the child could name the skill in at least one episode except shown. Q-values correspond to False Discovery Rate (FDR)-adjusted significance levels across all ten mechanisms, following Benjamini et al. (2006). Cluster-robust standard errors at the school level in parenthesis. Significance: *** p<0.01, ** p<0.05, * p<0.1

The show was produced with the goal of changing children's mindsets regarding their

[23]We further explore heterogeneity by language spoken at home in Appendix Table A17. We do not find evidence that the treatment differentially shifted children's reading preferences or habits by language spoken at home.

attitudes towards reading through creating characters that were relatable role models to children. Prior research has identified role model imitation as a crucial driver in media-based interventions (e.g., La Ferrara, 2016; Ahn et al., 2020; Greitemeyer, 2022; Riley, 2024), so why do preferences not change in this context? One possibility is that changes in preferences may depend on the extent to which children relate to on-screen characters. Thus, the absence of preference shifts reflect that children did not strongly identify with the main characters. Drawing on social learning theory (Bandura and Walters, 1977), that emphasizes the role of relatable models in shaping norms and behaviors, we measured character identification among treatment children. More than 50% of children reported that Nuzo, Namia, and Bubelang were *"Not at all like me,"* suggesting limited personal identification with the characters.

Second, the treatment could have induced children to change their time use – potentially crowding-in (or out) activities that contribute to learning. While studying compliance, we found there was a significant shift in TV watching habits among treated children towards the days Nuzo & Namia aired. In columns (4) to (6) of Table 6 we test whether the treatment induced in overall changes in child-reported watching TV more than 2 hours a day, parent-reported number of days a week their child watches TV, and whether child reports mostly watching non-educational television. We do not find evidence that the treatment significantly changed time children devoted to watching TV - either reported by the child or the parent. However, the treatment seems to have induced treated children to watch less non-educational TV shows by 4.7 percentage points (remains significant after adjusting for multiple testing, q-value$<$0.016). We also find no significant evidence that the treatment induced changes in the activities children engaged in on days the show was not aired (Appendix Table A18. Taken together, these results suggest that the encouragement treatment did not significantly change children's time use and habits as a result of the encouragement treatment.

Third, we examine the extent to which the information/skill content delivered by Nuzo & Namia could have changed outcomes. Unlike other educational TV interventions conveying specific information (e.g., on HIV prevention), Nuzo & Namia models key reading comprehension strategies by explicitly explaining them while simultaneously demonstrating their application in problem-solving contexts. One approach, as suggested by La Ferrara (2016), is testing whether there are differential gains by baseline levels of knowledge (or achievement in this context). However, Section 3.3 suggests that there is no significant differences by baseline levels of outcomes.[24] As detailed in the pre-registration, we attempt to obtain a proxy measure of changes in information in our compliance tool. At endline, during the episode

[24]Columns (31) to (33) of Appendix Table A5 suggest that the point estimates for comprehension is slightly larger for children with lower comprehension at baseline, though the difference is small and not statistically relevant. Similarly, SEL gains are larger for those with lower baseline levels, but again the differences are not statistically significant.

recognition task showing key moments employing different learning strategies, children were asked to describe the skill being learned. Thus, we can directly proxy for the informational component of the show.

The treatment significantly increased skill recognition. Metacognitive awareness of specific reading skills – understanding how to perform a strategy to enhance comprehension – is a cognitively demanding process at this developmental stage. Therefore, it is notable that treatment children are 10.6 percentage points more likely than control children to recognize and accurately explain at least one of the taught strategies (p-value and q-value < 0.01; see Column (7) in Table 6). This finding supports the interpretation that *information provision* serves as a key mechanism through which the edutainment intervention enhances reading comprehension.

4.3 Other Mechanisms

Beyond the potential barriers to compliance and the mechanisms explored, our encouragement might have also affected outcomes through indirect channels, unrelated to the TV show. For example, the initial information session along with SMS reminders could have increased parental investment in children's literacy or induced by changes in interactions with others (e.g., Di Maro et al., 2024; Duflo et al., 2011). To assess these mechanisms, we measured parental engagement using two survey items[25] and construct a social interaction index from five survey items.[26] Table 7 presents estimates of the treatment on these potential indirect channels. Overall, the point estimates for the three outcomes are close to zero, and we can reject economically small effects. These findings suggest that other factors like parental involvement or peer interactions did not play a meaningful role in driving changes in outcomes.[27]

5 Discussion

We document three main findings. First, while there is suggestive evidence of average gains in reading comprehension, we find meaningful heterogeneity based on whether English is spoken at home. Second, the treatment significantly increases curiosity. Third, while the

[25] How often does your mum, dad or carer *encourage you* to read? How often does your mum, dad or carer *read books or stories to you*?

[26] Social interaction index is based on responses to: how often children report on a four-point Likert scale reading with friends, discussing feelings with parents or peers, and seeking help from friends or parents when solving a problem.

[27] Also, note that the coefficient for "remedial classes" which could be a direct proxy for other parental investments in education is close to zero and not statistically relevant (Appendix Table A18).

Table 7: ITT Effects - Indirect Channels

Mechanism:	(1)	(2) Indirect Pathways	(3)
Dep. Var.:	Parent Encourage	Parents Read Stories	Social Interaction Index
Treatment	0.014	-0.012	0.016
	(0.012)	(0.020)	(0.026)
q-value	0.547	0.557	0.557
Control Mean	0.83	0.57	0.00
School Controls	Yes	Yes	Yes
Adj. R^2	0.013	0.016	0.042
Observations	3,669	3,662	3,669

Notes: Each column reports OLS regression estimates for a different outcome or specification. All regressions include quadruplet fixed effects and school controls. School-level controls (averages): share of girls, asset score, caregiver education, grade, reading score, HH size, TV watching time, and school size. The social interaction index is constructed following Anderson (2008). All other mechanism variables are binary indicators. Q-values correspond to False Discovery Rate (FDR)-adjusted significance levels across all ten mechanisms, following Benjamini et al. (2006). Cluster-robust standard errors at the school level in parenthesis. Significance: *** p<0.01, ** p<0.05, * p<0.1

show sought to challenge traditional gender norms, we observe a robust backlash effect, where the treatment reinforce traditional gender roles. In this section, we explore potential explanations and implications of these findings.

We find that watching Nuzo & Namia appears to improve comprehension in our sample on average; however, this effect is largely driven by gains among children who predominantly speak English at home, with no significant effects observed among non-English-speaking children at home. This result aligns closely with evidence highlighting the importance of language congruence between instructional materials and children's home languages. For example, evidence from neuroscience suggests that early exposure to language at home strengthens neural pathways between the key language processing regions of the brain (Romeo et al., 2018). Importantly, Mondt et al. (2011) find that when bilingual children attempt to comprehend instructions and solve arithmetic problems in their school language, there was reduced activation in key brain regions compared to when the tasks were conducted in their home language. Furthermore, there is extensive evidence in education and psychology documenting the importance of language congruence between a child's home language and language of instruction for learning (Bermejo et al., 2021; Carlo et al., 2014; Malakoff, 1988).

The limited effectiveness of the show for children not speaking English at home raises the

question about whether foundational comprehension skills acquired in one language readily transfer to another. Prior research suggests that skill transfer between languages may be constrained, particularly when foundational language proficiency is weak or underdeveloped (Malakoff, 1988; Cummins, 1979). To directly test this, Baier et al. (2025) conducted a controlled, small-scale lab-in-the-field experiment following our main field experiment. The study examined the effectiveness of Nuzo & Namia in Kiswahili among children in Kenya on the same outcomes. However, the literacy assessments were administered in three languages: Kiswahili, English, and Luo. They find that watching the show in a controlled environment increased comprehension in Kiswahili (and the SEL index) by 42.4% (and 34.5%) of a standard deviation. However, there were no significant effects on English or Luo comprehension, suggesting that comprehension skills may not readily transfer across languages. Note that the estimated effects in Baier et al. (2025) are consistent with the LATE estimates from our field experiment, suggesting that a show that is effective in a small-scale and controlled environment can indeed generate similar benefits in a uncontrolled real-world environment.

Evidence from our study indicates that children exposed to Nuzo & Namia exhibit a significant increase in curiosity. Prior research on interventions to boost curiosity in children has produced mixed results. For example, a recent meta-analysis by Schutte and Malouff (2023) shows that research in this area is limited: only 4 out of 41 studies focus on children, and the average effect size across studies is modest (about 33% of a standard deviation) and not statistically significant. Notably, these studies typically featured sample sizes of fewer than 100 participants. By contrast, our findings, together with those of Alan and Mumcu (2024), demonstrate that large-scale interventions can meaningfully foster curiosity. This raises the question of what kinds of growth and learning greater curiosity may foster. Although the program intended to channel additional motivation into increased reading behavior, our mechanism analysis does not detect a corresponding rise in reported reading. Therefore, it is plausible that the intervention fostered a more general form of epistemic curiosity (Berlyne, 1954) – reflecting a stable personal trait rather than a transient state (Litman and Spielberger, 2003). Given that previous research links higher curiosity with improved attention and retention (Kang et al., 2009; Wojtowicz and Loewenstein, 2020; Alan and Mumcu, 2024), it is possible that the program's impact on curiosity also contributed to the observed comprehension gains. In this view, curiosity may have enhanced skill recognition and memory retention. However, our analysis is not able to disentangle these effects from those of direct instruction.

Our results suggest that strong norm-messaging may produce counterproductive effects. Prior studies demonstrate that TV and film interventions can reduce fertility in Brazil and the U.S. (Ferrara et al., 2012; Kearney and Levine, 2015), and that films featuring female

role models can improve educational outcomes for girls in Uganda (Riley, 2024). However, while these studies underscore the importance of role models, they do not systematically consider the critical role of relatability. According to the motivational theory of role modeling (Morgenroth et al., 2015), the effectiveness of gender non-conforming role models in shifting children's attitudes and behavior hinges on the extent to which children can identify with them.

To test this theory, we conducted a supplementary analysis, outside the original pre-analysis plan, to shed light on the role of character relatability. Because treatment did not affect which characters children found relatable, we estimate equation 1, interacting the treatment with three indicator variables for whether the child relates to each of the main characters. The estimates presented in Appendix Table A19 suggest that the negative gender role effects are concentrated among children who did not relate to the character Namia, while such heterogeneity is absent for other characters. Namia, the sister, was the primary character who exhibited gender nonconformity and conveyed nontraditional gender norms. She was also portrayed as neurodivergent, with her autism diagnosis serving as a recurring theme throughout the show.[28] It is possible that the combined portrayal of neurodivergence and gender nonconformity in Namia was too removed from the children's everyday experiences. Indeed, there is evidence of significant stigma regarding autism among children. For example, children rated peers labeled with autism less favorably (Swaim and Morgan, 2001), perceived them as less similar to themselves (Harnum et al., 2007), and showed a preference for peers without disabilities (Nowicki and Sandieson, 2002). These findings suggest that improving character relatability, perhaps by more explicitly emphasizing shared traits even within a gender non-conforming framework as suggested by Olsson and Martiny (2018), could potentially reduce the documented backfiring effect.

6 Estimating Cost-effectiveness

We found that an edutainment intervention can be effective in increasing targeted skills, namely comprehension and socio-emotional learning, in a naturalistic setting. Much of the appeal surrounding educational television and educational technology (Ed-Tech) emphasizes their low cost and scalability. We assess this quantitatively and show that even at scale this intervention compares favorably to some of the most cost-effective education interventions, such as deworming, and is more cost-effective than most structured pedagogy programs (Angrist et al., 2025; Miguel and Kremer, 2004). We report cost-effectiveness based solely on the comprehension outcome, given the interpretive ambiguity surrounding SEL and norm-

[28] See https://www.youtube.com/watch?v=f7Uz1f2VVvg for an excerpt.

related measures.

6.1 Methodology

The first two rows of Table 8 shows annualized intervention costs, including costs borne by households, calculated following J-PAL's ingredient method discussed in McEwan (2012) and Dhaliwal et al. (2013). To make estimates of cost-effectiveness easily comparable with previous studies, we present results in inflation-adjusted US dollars for both the analysis year (2024) and for 2011. Ubongo, the show's producer, provided raw data from its financial and project management systems related to the production and distribution of the educational TV program. We compiled and cross-verified over 5,600 cost entries from these records, organizing them into 18 distinct categories (see Table A20, showing full investment costs).[29]

We estimated at-scale encouragement, beneficiary, and distribution costs[30] using internal records, monthly average domestic tariff rates from Kenya Power and Lighting Company (KPLC), and Akili Kids' broadcasting rate card, respectively. Here, "at scale" refers to multiplying the study's per-child cost estimates by the total number of children reached on the Kenyan TV channel that aired Nuzo & Namia. Geopoll data from July and December 2022 indicate that Akili Kids reaches an average of 1,238,000 children during Ubongo's broadcast times. Our analysis focuses solely on the Kenyan audience to align with the population for which impacts were measured. However, because the program was produced for and aired in five countries, the true production costs per child are lower, rendering our cost-effectiveness estimates conservative. Assuming that the comprehension and social-emotional learning gains observed can be achieved across all markets, cost-effectiveness would be even more favorable, as distribution costs increase only marginally with additional market exposure.

In the context of a television program as an intervention, we classify the production of the show as an investment cost. To determine the expected relevance and longevity of such programs, we refer to the US and UK's Generally Accepted Accounting Principles (GAAP), as well as historical evidence from similar shows, including Sesame Street, Dora the Explorer, and Ubongo Kids. Our calculations suggest that each season of a program of this nature is likely to remain relevant and be broadcast in reruns for approximately ten years. Consequently, we allocate one-tenth of the investment cost in our annual cost estimate. We incorporate non-investment-related costs in our annual cost estimate to their full extent,

[29] We used task descriptions from Ubongo's internal project management system, along with financial data, to organize all production and implementation costs. Two researchers completed task and cost categorization separately and subsequently discussed any disagreements to ensure unbiased classification. As pre-registered, we use a 10% discount rate and make inflation adjustments using the GDP deflator.

[30] Encouragement costs encompass parent outreach and SMS reminders; beneficiary costs reflect estimated electricity expenses for television use; and distribution costs pertain to channel airing fees.

including Ubongo's organizational overheads, encouragement and targeting expenses, channel airing fees, and costs borne by beneficiaries over the course of the year-long intervention.

6.2 Cost-effectiveness estimates

Following Angrist et al. (2025) we calculate a quality adjusted measure of impact to allow for cross-country comparison of learning interventions. We find that at 6.77 learning adjusted years of schooling (LAYS) per $100 USD for comprehension increases in a naturalistic at-scale setting, the intervention is more cost-effective than 91% of the 135 education interventions with cost-effectiveness data included in Angrist et al. (2025).

Table 8: Cost-effectiveness estimates

	(1) All Children	(2) English Speaking
Outcome:	Comprehension	
Annual Costs (2024 real USD)	$2,059,142.24	$479,977.42
Annual Costs (2011 real USD)	$1,504,307.47	$350,647.77
Total Benefits Accrued (Std.Dev.)	111,420	51,450
Benefit per Child (Std.Dev.)	0.090	0.230
Cost per Child (2024 real USD)	$1.66	$2.15
Cost per Child (2011 real USD)	$1.22	$1.57
Cost per Std. Dev. (2024 real USD)	$18.48	$9.33
Cost per Std. Dev. (2011 real USD)	$13.50	$6.82
LAYS per $100	6.77	13.37

Notes: Annual costs comprise at-scale encouragement, beneficiary, and distribution costs, as well as partial investments in content production as detailed in Section 6.1. To represent benefits, we use coefficients for comprehension from our primary ITT regressions and ITT regressions studying heterogeneity (Panel A Column (4) in Table 5, and column (1) of Appendix Table A5). Gains as well as encouragement and beneficiary costs are scaled by the total estimated Kenyan audience (see Section 6.1). The at-scale audience in Column (2) of this table was reduced proportionally to 223,696 using the share of english speaking households with functioning TV in our endline survey.

We also compute cost-per-child estimates and cost-per-standard deviation gains to contextualize our findings, noting that these present conservative estimates given that we do not consider the full scale for which this content was produced (see Section 6.1). Venetis, Emma and Harris-Van Keuren, Christine (2024) and Harris-Van Keuren, Christine (2023) report that the costs of education interventions vary widely, ranging from $1.64 to $128.07 per student. With a cost-per-child of $1.66 (or $1.41 when adjusted to 2020 USD), this intervention is positioned at the very low end of this cost distribution. It compares favorably

to other education programs, such as a structured pedagogy program ($7.75, 2020 USD) and Teaching at the Right Level (TaRL) ($19.53, 2020 USD, without technology use, and $26.64 with technology use) (Angrist et al., 2023). Even if the true effects were smaller than our ITT estimates, the intervention remains highly cost-effective given its low cost-per-child. Compared with a structured pedagogy program that costs $59.62 per standard deviation gain, even learning improvements of just 0.028 standard deviations would yield comparable cost-effectiveness. These figures are considerably lower than the $130.20 and $96.87 per standard deviation gain (2020 USD, without and with technology use, respectively) reported for Teaching at the Right Level (Angrist et al., 2023).

7 Conclusions

Using a randomized encouragement design, we provide evidence that naturalistic, unguided exposure to over-the-air educational television improves children's comprehension skills and curiosity, while also reinforcing more traditional gender role attitudes. These effects are largest among children who predominantly speak English at home. Analysis of potential mechanisms indicates that the impacts are primarily driven by the informational and skill-building content of the programming, rather than changes in preferences or time use. Taken together, our findings demonstrate that educational television interventions that are effective in controlled settings can scale successfully through existing broadcast infrastructure, and deliver meaningful benefits at low cost. However, because Kiswahili is the predominant language spoken by children in our sample, the gains from the show are not equally distributed and, as a result, may exacerbate existing language achievement gaps.

While we study the effects of an educational TV show in Kenya, there are several features of the intervention and context that can translate to other countries. First, the television show was produced to appeal to children across Sub-Saharan Africa, by a producer with experience in creating content for the region. Second, the broadcast medium (free over-the-air in a home setting) is common. Third, many countries in the region face similar early-exit language models at school where the skills conveyed by the show are particularly relevant. However, there are also some limitations to external validity. For instance, the specific language mix in the household or cultural attitudes towards gender could attenuate the effectiveness of the show. An additional constraint on scalability is that access to TV remains limited, particularly in rural or poorer regions. However, this constraint is likely to become less binding over time, considering the growing penetration of television and projected improvements in access to electricity across the region (World Bank, 2024).

This study opens up multiple avenues for future research. First, future work could exam-

ine long-run effects from improving comprehension and confidence on performance in other subjects. Second, while we find suggestive evidence that character relatability may play a role in the observed gender role backlash, future work could explore which specific character attributes contribute to this effect. Third, an aspect that remains unexplored about Nuzo & Namia is the extent to which it is able to foster a sense of Pan-African identity and intercultural understanding. This may have broader implications for fostering social cohesion and mitigating inter-ethnic tensions.

References

Abeberese, A. B., T. J. Kumler, and L. L. Linden (2014). Improving reading skills by encouraging children to read in school:: A randomized evaluation of the sa aklat sisikat reading program in the philippines. *Journal of Human Resources 49*(3), 611–633.

Ahn, J. N., D. Hu, and M. Vega (2020). "Do as I do, not as I say": Using social learning theory to unpack the impact of role models on students' outcomes in education. *Social and Personality Psychology Compass 14*(2), e12517.

Alan, S., T. Boneva, and S. Ertac (2019). Ever failed, try again, succeed better: Results from a randomized educational intervention on grit. *The Quarterly Journal of Economics 134*(3), 1121–1162.

Alan, S. and S. Ertac (2018). Fostering patience in the classroom: Results from randomized educational intervention. *Journal of Political Economy 126*(5), 1865–1911.

Alan, S. and I. Mumcu (2024). Nurturing childhood curiosity to enhance learning: Evidence from a randomized pedagogical intervention. *American Economic Review 114*(4), 1173–1210.

Algan, Y., E. Beasley, S. Côté, J. Park, R. E. Tremblay, and F. Vitaro (2022). The impact of childhood social skills and self-control training on economic and noneconomic outcomes: Evidence from a randomized experiment using administrative data. *American Economic Review 112*(8), 2553–2579.

Anderson, M. L. (2008). Multiple inference and gender differences in the effects of early intervention: A reevaluation of the abecedarian, perry preschool, and early training projects. *Journal of the American Statistical Association 103*(484), 1481–1495.

Angrist, N., E. Aurino, H. A. Patrinos, G. Psacharopoulos, E. Vegas, R. Nordjo, and B. Wong (2023). Improving Learning in Low- and Lower-Middle-Income Countries. *Journal of Benefit-Cost Analysis 14*(S1), 55–80.

Angrist, N., D. K. Evans, D. Filmer, R. Glennerster, H. Rogers, and S. Sabarwal (2025). How to improve education outcomes most efficiently? A review of the evidence using a unified metric. *Journal of Development Economics 172*, 103382.

Araya, R., E. Arias Ortiz, N. L. Bottan, and J. Cristia (2024). Integrating learning platforms within regular school time: Experimental evidence from chilean primary schools.

Aulagnon, R., N. Bottan, J. Cristia, and M. Romero (2025). From static to adaptive: Comparing the efficacy of messaging strategies in student online engagement. Working Paper.

Baier, J., N. Bottan, and A. Ghosh (2025). Investigating learning mechanisms for nuzo and namia show. *mimeo*.

Bandura, A. and R. H. Walters (1977). *Social learning theory*, Volume 1. Englewood cliffs Prentice Hall.

Banerjee, A., R. Banerji, J. Berry, E. Duflo, H. Kannan, S. Mukerji, M. Shotland, and M. Walton (2017). From proof of concept to scalable policies: Challenges and solutions, with an application. *Journal of Economic Perspectives 31*(4), 73–102.

Banerjee, A., E. La Ferrara, and V. H. Orozco-Olvera (2019). The entertaining way to behavioral change: Fighting hiv with mtv. Technical report, National Bureau of Economic Research.

Benjamini, Y., A. M. Krieger, and D. Yekutieli (2006). Adaptive linear step-up procedures that control the false discovery rate. *Biometrika 93*(3), 491–507.

Berlyne, D. E. (1954). A theory of human curiosity. *British Journal of Psychology. General Section 45*(3), 180–191.

Bermejo, V., P. Ester, and I. Morales (2021). How the language of instruction influences mathematical thinking development in the first years of bilingual schoolers. *Frontiers in Psychology 12*, 533141.

Borzekowski, D. L. (2018). A quasi-experiment examining the impact of educational cartoons on tanzanian children. *Journal of Applied Developmental Psychology 54*, 53–59.

Bruhn, M. and D. McKenzie (2009). In pursuit of balance: Randomization in practice in development field experiments. *American Economic Journal: Applied Economics 1*(4), 200–232.

Cardim, J., T. Molina-Millán, and P. C. Vicente (2023). Can technology improve the classroom experience in primary education? an african experiment on a worldwide program. *Journal of Development Economics 164*, 103145.

Carlo, M. S., C. D. Barr, D. August, M. Calderón, and L. Artzi (2014). Language of instruction as a moderator for transfer of reading comprehension skills among spanish-speaking english language learners. *Bilingual Research Journal 37*(3), 287–310.

Cristia, J. P., S. Cueto, O. Malamud, and R. Aulagnon (2024). Streaking to success: The effects of highlighting streaks on student effort and achievement.

Cummins, J. (1979). Linguistic interdependence and the educational development of bilingual children. *Review of Educational Research 49*(2), 222–251.

De Quidt, J., J. Haushofer, and C. Roth (2018). Measuring and bounding experimenter demand. *American Economic Review 108*(11), 3266–3302.

De Quidt, J., L. Vesterlund, and A. J. Wilson (2019). Experimenter demand effects. *Handbook of Research Methods and Applications in Experimental Economics 384*.

Dhaliwal, I., E. Duflo, R. Glennerster, and C. Tulloch (2013). Comparative cost-effectiveness analysis to inform policy in developing countries: A general framework with applications for education. *Education Policy in Developing Countries*.

Di Maro, V., S. Leeffers, D. Serra, and P. C. Vicente (2024). Mobilizing parents at home and at school: an experiment on primary education in angola. *Economic Development and Cultural Change 72*(3), 1069–1116.

Duflo, E., P. Dupas, and M. Kremer (2011). Peer effects, teacher incentives, and the impact of tracking: Evidence from a randomized evaluation in kenya. *American Economic Review 101*(5), 1739–1774.

Durlak, J. A., R. P. Weissberg, A. B. Dymnicki, R. D. Taylor, and K. B. Schellinger (2011). The Impact of Enhancing Students' Social and Emotional Learning: A Meta-Analysis of School-Based Universal Interventions. *Child Development 82*(1), 405–432.

Eberhard, D. M., G. F. Simons, and C. D. Fennig (Eds.) (2025). *Ethnologue: Languages of the World* (Twenty-eighth edition ed.). Dallas, Texas: SIL International.

Ferrara, E. L., A. Chong, and S. Duryea (2012). Soap operas and fertility: Evidence from brazil. *American Economic Journal: Applied Economics 4*(4), 1–31.

Fryer Jr, R. G. (2011). Financial incentives and student achievement: Evidence from randomized trials. *The Quarterly Journal of Economics 126*(4), 1755–1798.

Glewwe, P. and K. Muralidharan (2016). Improving education outcomes in developing countries: Evidence, knowledge gaps, and policy implications. In *Handbook of the Economics of Education*, Volume 5, pp. 653–743. Elsevier.

Graham, J. and S. Kelly (2019). How effective are early grade reading interventions? a review of the evidence. *Educational Research Review 27*, 155–175.

Greitemeyer, T. (2022). Prosocial modeling: person role models and the media. *Current Opinion in Psychology 44*, 135–139.

Harnum, M., J. Duffy, and D. A. Ferguson (2007). Adults' versus children's perceptions of a child with autism or attention deficit hyperactivity disorder. *Journal of autism and developmental disorders 37*, 1337–1343.

Harris-Van Keuren, Christine (2023). Learning at scale: Time and money. Technical report, RTI International.

Hungi, N. and F. W. Thuku (2010). Differences in pupil achievement in kenya: Implications for policy and practice. *International Journal of Educational Development 30*(1), 33–43.

Jensen, R. and E. Oster (2009). The power of tv: Cable television and women's status in india. *The Quarterly Journal of Economics 124*(3), 1057–1094.

Jukes, M. C. H., N. L. Mgonda, J. J. Tibenda, P. Gabrieli, G. Jeremiah, K. L. Betts, J. Williams, and K. L. Bub (2021). Building an assessment of community-defined social-emotional competencies from the ground up in tanzania. *Child Development 92*, 1095–1109.

Kang, M. J., M. Hsu, I. M. Krajbich, G. Loewenstein, S. M. McClure, J. T.-y. Wang, and C. F. Camerer (2009). The Wick in the Candle of Learning: Epistemic Curiosity Activates Reward Circuitry and Enhances Memory. *Psychological Science 20*(8), 963–973.

Kashdan, T. B., F. R. Goodman, D. J. Disabato, P. E. McKnight, K. Kelso, and C. Naughton (2020). Curiosity has comprehensive benefits in the workplace: Developing and validating a multidimensional workplace curiosity scale in United States and German employees. *Personality and Individual Differences 155*, 109717.

Kearney, M. S. and P. B. Levine (2015). Media influences on social outcomes: The impact of mtv's 16 and pregnant on teen childbearing. *American Economic Review 105*(12), 3597–3632.

Kearney, M. S. and P. B. Levine (2019). Early childhood education by television: Lessons from sesame street. *American Economic Journal: Applied Economics 11*(1), 318–350.

Keaveney, E., C. Fierros, A. Rigaux, and A. Menendez (2021). Tusome external evaluation end line report.

Kenya Ministry of Education (2019). Basic education statistical booklet. Technical report, Ministry of Education.

Kenya Ministry of Education (2020). Basic education statistical booklet. Technical report, Ministry of Education.

Kerwin, J. T. and R. L. Thornton (2021). Making the grade: The sensitivity of education program effectiveness to input choices and outcome measures. *Review of Economics and Statistics 103*(2), 251–264.

KICD (2025). Series 1-Curriculum Reforms in Kenya. Technical report, Ministry of Education.

Kwan, K. M. W., S. Y. Shi, A. N. Nabbijohn, L. N. MacMullin, D. P. VanderLaan, and W. I. Wong (2020). Children's Appraisals of Gender Nonconformity: Developmental Pattern and Intervention. *Child Development 91*(4).

La Ferrara, E. (2016). Mass media and social change: Can we use television to fight poverty? *Journal of the European Economic Association 14*(4), 791–827.

Lai, F., R. Luo, L. Zhang, X. Huang, and S. Rozelle (2015). Does computer-assisted learning improve learning outcomes? evidence from a randomized experiment in migrant schools in beijing. *Economics of Education Review 47*, 34–48.

Litman, J. A. and C. D. Spielberger (2003). Measuring Epistemic Curiosity and Its Diversive and Specific Components. *Journal of Personality Assessment 80*(1), 75–86.

Lydon-Staley, D. M., P. Zurn, and D. S. Bassett (2020). Within-person variability in curiosity during daily life and associations with well-being. *Journal of Personality 88*(4), 625–641.

Malakoff, M. E. (1988). The effect of language of instruction on reasoning in bilingual children. *Applied Psycholinguistics 9*(1), 17–38.

Mares, M.-L. and Z. Pan (2013). Effects of sesame street: A meta-analysis of children's learning in 15 countries. *Journal of Applied Developmental Psychology 34*(3), 140–151.

McEwan, P. J. (2012). Cost-effectiveness analysis of education and health interventions in developing countries. *Journal of Development Effectiveness 4*(2), 189–213.

McKenzie, D. (2022). Why i am now more cautious about using or recommending matched pair randomization and like matched quadruplets instead.

Miguel, E. and M. Kremer (2004). Worms: identifying impacts on education and health in the presence of treatment externalities. *Econometrica 72*(1), 159–217.

Mondt, K., E. Struys, M. Van Den Noort, D. Balériaux, T. Metens, P. Paquier, P. Van De Craen, P. Bosch, and V. Denolin (2011). Neural Differences in Bilingual Children's Arithmetic Processing Depending on Language of Instruction. *Mind, Brain, and Education 5*(2), 79–88.

Moran, C., D. Hilgendorf, V. Y. Zhao, D. Al-Ogaily, H. Yoshikawa, K. Schwartz, J. Rafla, A. Molano, K. Strouf, M. Khanji, R. A. Seriah, M. A. Aabed, R. Fityan, P. Sloane, L. Hussein, D. Hidayah, M. Shukri, T. Sharawi, K. Foulds, A. B. Bowden, S. Lee, P. Tiwari, and J. Behrman (2023). Lessons and Impacts of Ahlan Simsim TV Program in Pre-Primary Classrooms in Jordan on Children's Emotional Development: A Randomized Controlled Trial.

Morgenroth, T., M. K. Ryan, and K. Peters (2015). The Motivational Theory of Role Modeling: How Role Models Influence Role Aspirants' Goals. *Review of General Psychology 19*(4), 465–483.

Muralidharan, K. and P. Niehaus (2017). Experimentation at scale. *Journal of Economic Perspectives 31*(4), 103–24.

Muralidharan, K., A. Singh, and A. J. Ganimian (2019). Disrupting education? experimental evidence on technology-aided instruction in india. *American Economic Review 109*(4), 1426–1460.

Murunga, J. (2021). Radio, tv and internet audience statistics in kenya – q1 2021.

Mussel, P. (2013). Introducing the construct curiosity for predicting job performance. *Journal of Organizational Behavior 34*(4), 453–472.

Nowicki, E. A. and R. Sandieson (2002). A meta-analysis of school-age children's attitudes towards persons with physical or intellectual disabilities. *International Journal of Disability, Development and Education 49*(3), 243–265.

Ogembo, J. G., B. Ngugi, and M. Pelowski (2012). Computerizing primary schools in rural kenya: Outstanding challenges and possible solutions. *The Electronic Journal of Information Systems in Developing Countries 52*(1), 1–17.

Olsson, M. and S. E. Martiny (2018). Does Exposure to Counterstereotypical Role Models Influence Girls' and Women's Gender Stereotypes and Career Choices? A Review of Social Psychological Research. *Frontiers in Psychology 9*, 2264.

Riley, E. (2024). Role models in movies: the impact of queen of katwe on students' educational attainment. *Review of Economics and Statistics 106*(2), 334–351.

Romeo, R. R., J. Segaran, J. A. Leonard, S. T. Robinson, M. R. West, A. P. Mackey, A. Yendiki, M. L. Rowe, and J. D. Gabrieli (2018). Language exposure relates to structural neural connectivity in childhood. *Journal of Neuroscience 38*(36), 7870–7877.

Schutte, N. S. and J. M. Malouff (2023). A meta-analytic investigation of the impact of curiosity-enhancing interventions. *Current Psychology 42*(24), 20374–20384.

Schwab, B., S. Janzen, N. P. Magnan, and W. M. Thompson (2020). Constructing a summary index using the standardized inverse-covariance weighted average of indicators. *The Stata Journal 20*(4), 952–964.

Semenova, V. and V. Chernozhukov (2021). Debiased machine learning of conditional average treatment effects and other causal functions. *The Econometrics Journal 24*(2), 264–289.

Swaim, K. F. and S. B. Morgan (2001). Children's attitudes and behavioral intentions toward a peer with autistic behaviors: does a brief educational intervention have an effect? *Journal of autism and developmental disorders 31*, 195–205.

Taylor, R. D., E. Oberle, J. A. Durlak, and R. P. Weissberg (2017). Promoting Positive Youth Development Through School-Based Social and Emotional Learning Interventions: A Meta-Analysis of Follow-Up Effects. *Child Development 88*(4), 1156–1171.

UNESCO (2025). *Languages matter: global guidance on multilingual education.* UNESCO.

UNICEF (2022). The state of global learning poverty: 2022 update.

USAID (2016). Early grade reading assessment (egra) toolkit. Technical report.

Venetis, Emma and Harris-Van Keuren, Christine (2024). Cost across contexts: A cross-country cost analysis of seven usaid-funded education programs. Technical report, USAID.

Wager, S. and S. Athey (2018). Estimation and inference of heterogeneous treatment effects using random forests. *Journal of the American Statistical Association 113*(523), 1228–1242.

Wojtowicz, Z. and G. Loewenstein (2020). Curiosity and the economics of attention. *Current Opinion in Behavioral Sciences 35*, 135–140.

Wooldridge, J. M. (2015). Control function methods in applied econometrics. *Journal of Human Resources 50*(2), 420–445.

World Bank (2024). Energizing africa - overview. Accessed: 2025-03-18.

Yadlowsky, S., S. Fleming, N. Shah, E. Brunskill, and S. Wager (2024). Evaluating treatment prioritization rules via rank-weighted average treatment effects. *Journal of the American Statistical Association*, 1–14.

York, B. N., S. Loeb, and C. Doss (2019). One step at a time: The effects of an early literacy text-messaging program for parents of preschoolers. *Journal of Human Resources 54*(3), 537–566.

Online Appendix (For Online Publication Only)

Baier et al., "Learning Through Living Room Screens"

March 19, 2025

A Appendix Tables and Figures

Table A1: Baseline Characteristics comparing Attritors and Non-Attritors

	(1)	(2)	(3)	(4)	(5)	(6)
	Treatment (Mean)		Control (Mean)		p-value	
	Non-Attritors	Attritors	Non-Attritors	Attritors	Non-Attritors	Attritors
Demographics:						
Child is female	0.505	0.526	0.501	0.497	0.397	0.291
Child in Grade 1	0.300	0.313	0.313	0.322	0.074	0.471
Child in Grade 2	0.352	0.361	0.329	0.345	0.067	0.798
Child in Grade 3	0.348	0.325	0.358	0.333	0.137	0.726
Hours watch TV	1.433	1.413	1.451	1.391	0.162	0.571
Watch 2+ hours	0.360	0.341	0.379	0.363	0.451	0.205
Watch Akili Kids Channel	0.442	0.441	0.432	0.423	0.773	0.805
HH size	5.619	5.479	5.603	5.495	0.378	0.175
HH has TV	0.924	0.924	0.904	0.898	0.366	0.757
HH has smartphone	0.593	0.594	0.564	0.563	0.424	0.212
HH Asset score	4.044	3.888	3.951	4.003	0.185	0.660
Caregiver has primary or less	0.625	0.609	0.643	0.610	0.677	0.934
HH speaks english	0.226	0.188	0.178	0.217	0.074	0.538
Nr books	14.504	11.529	14.252	12.231	0.839	0.344
Baseline Outcomes:						
Literacy Index	0.025	-0.024	-0.018	0.090	0.772	0.758
Comprehension Index	0.009	-0.064	-0.011	0.057	0.554	0.348
Reading Index	0.033	-0.000	-0.013	0.066	0.810	0.685
SEL Index	0.047	-0.003	-0.011	0.055	0.097	0.729
Confidence score (z)	0.047	0.054	-0.000	0.025	0.154	0.485
Curiosity score (z)	0.054	-0.038	-0.000	0.089	0.128	0.204
Gender Attitude Index	0.001	-0.063	0.004	-0.017	0.698	0.383
Gender stereotype (z)	-0.041	-0.040	0.014	-0.070	0.191	0.769
Gender roles (z)	0.027	-0.023	-0.001	0.004	0.623	0.719
Gender bias (z)	0.018	-0.128	-0.008	0.039	0.753	0.072
Observations	1858	340	1811	364		

Notes: Columns (1) to (4) present means for baseline characteristics for attritors and non-attritors in the treatment and control groups. Columns (5) and (6) present p-values for testing the regression adjusted difference in means between treatment and control for non-attritors (column (5)) and attritors (column (6)). OLS regressions include quadruple fixed effects and school-level controls. School-level controls (averages): share of girls, asset score, caregiver education, grade, reading score, HH size, TV watching time, and school size. Cluster-robust standard errors were used.

Table A2: Benchmarking EGRA Scores with a Kenyan Nationally Representative Survey

	(1)	(2)	(3)	(4)	(5)	(6)	(7)
		Baseline Survey			Kenya 2021 (USAID)		
	Overall	Grade 1	Grade 2	Grade 3	Grade 1	Grade 2	Grade 3
CWPM							
Letter Recognition	24.67	20.30	25.52	27.64	20.70	27.10	30.00
Non-word Recognition	7.96	4.15	7.30	12.02	7.80	14.80	24.10
Oral Fluency	40.23	15.24	34.10	68.14	11.40	27.80	55.70
% Correct							
Letter Recognition	24.64	20.25	25.51	27.60	20.70	27.00	29.90
Non-word Recognition	15.71	7.87	14.54	23.83	15.40	29.30	46.30
Oral Fluency	36.95	18.21	36.74	53.78	15.80	37.90	64.30
Reading Comprehension	33.30	15.63	32.47	50.07	8.60	21.90	42.70
Listening Comprehension	64.81	57.51	64.23	72.13	NA	NA	NA
% Zero score							
Letter Recognition	2.49	3.39	2.58	1.66	21.60	15.40	NA
Non-word Recognition	17.68	28.81	16.62	8.98	52.20	28.70	NA
Oral Fluency	24.79	39.22	25.10	11.70	45.00	26.60	NA
Reading Comprehension	47.79	71.64	47.35	26.93	80.00	53.90	NA
Listening Comprehension	10.95	16.67	10.65	6.05	NA	NA	NA

Notes: Correct words (or sounds) per minute (CWPM) calculated as (Items attempted - Items incorrect)/((duration-time remaining)/duration). Kenya 2021: USAID Kenya Tusome Early Grade Reading Activity Study on Grade 1, 2, and 3 Pupil Learning Outcomes in 2021: Final Report. URL: https://pdf.usaid.gov/pdf_docs/PA00ZXKJ.pdf [backup link]

Table A3: Effects of Encouragement on Compliance by Baseline Literacy

	(1)	(2)	(3)	(4)	(5)	
		Self-Report:			Recall:	
Dep. Var.:	Ever watch	Consistent watch	% Episode watch	Char. ID	Consistent watch (ID'd)	% Episode ID'd
Treatment	0.441***	0.171***	0.178***	0.321***	0.054***	0.075***
	(0.024)	(0.017)	(0.012)	(0.025)	(0.014)	(0.010)
Treat. X High Lit.	0.037	0.091***	0.050***	0.068**	0.064***	0.046***
	(0.028)	(0.022)	(0.016)	(0.030)	(0.018)	(0.014)
Mean Control-Low	0.28	0.06	0.08	0.08	0.03	0.04
Mean Control-High	0.34	0.06	0.10	0.14	0.04	0.05
School Controls	Yes	Yes	Yes	Yes	Yes	Yes
R^2	0.321	0.164	0.237	0.255	0.112	0.176
Observations	3,668	3,668	3,668	3,668	3,668	3,668

Notes: Each column presents estimates corresponding to an OLS regression for a different outcome by level of the baseline literacy index, where high literacy scored above the median. All estimates include quadruplet fixed effects and school-level controls – all interacted with indicator for high literacy. School-level controls (averages): share of girls, asset score, caregiver education, grade, reading score, HH size, TV watching time, and school size. Cluster-robust standard errors at the school level in parenthesis. Significance: *** p<0.01, ** p<0.05, * p<0.1

Table A4: ITT Estimates - Secondary Outcomes

	(1)	(2)	(3)	(4)
	Treatment			
	Coef.	S.E.	q-value	Obs.
Panel A: Literacy				
Comprehension Index	0.090*	(0.054)		3,669
Reading comprehension (z)	0.084*	(0.046)	0.349	3,669
Listening comprehension (z)	0.078	(0.055)	0.399	3,669
Reading Index	0.006	(0.046)		3,669
Non-word reading (z)	-0.015	(0.047)	0.789	3,663
Letter recognition (z)	-0.014	(0.051)	0.789	3,668
Oral fluency (z)	0.035	(0.035)	0.548	3,653
Panel B: Gender Attitudes				
Gender roles (z)	-0.110**	(0.043)		3,669
Girl Gender roles (z)	-0.072*	(0.040)	0.217	3,669
Boy Gender roles (z)	-0.065**	(0.032)	0.217	3,669
Gender bias (z)	-0.112	(0.087)		3,669
Boy Gender bias: willing to interact (z)	-0.007	(0.026)	0.796	1,822
Girl Gender bias: willing to interact (z)	-0.036	(0.031)	0.487	1,847
Boy Gender bias: interactions (z)	0.006	(0.024)	0.796	1,822
Girl Gender bias: interactions (z)	-0.021	(0.027)	0.669	1,847

Notes: Each row presents estimates for an OLS regression corresponding to a different outcome. Each panel presents components of primary indices for our three main outcomes. All estimates include controls for the baseline dependent variable, school controls and quadruplet fixed effects. School-level controls (averages): share of girls, asset score, caregiver education, grade, reading score, HH size, TV watching time, and school size. Cluster-robust standard errors at the school level in parenthesis. Q-values in column (4) correspond to False Discovery Rate (FDR)-adjusted significance levels within each family of outcomes, following Benjamini et al. (2006). Significance: *** p<0.01, ** p<0.05, * p<0.1

Table A5: ITT Effects - Heterogeneity

Heterogeneity Var.:	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)	(13)	(14)	(15)
	Language at home			Gender			Has books at home			Caregiver education			Household size		
	English	Other	p-value diff.	Boy	Girl	p-value diff.	Yes	No	p-value diff.	High	Low	p-value diff.	Large	Small	p-value diff.
Panel A: Dep. Vars. for Literacy															
Literacy Index	0.184**	0.010	0.031	0.059	0.045	0.775	0.036	0.068	0.606	0.063	0.052	0.854	0.044	0.061	0.792
	(0.083)	(0.053)		(0.058)	(0.056)		(0.062)	(0.056)		(0.062)	(0.056)		(0.070)	(0.053)	
Comprehension Index	0.230**	0.051	0.042	0.116*	0.069	0.341	0.092	0.099	0.921	0.102	0.088	0.815	0.082	0.096	0.829
	(0.094)$^{+}$	(0.054)		(0.061)	(0.058)		(0.060)	(0.062)		(0.066)	(0.057)		(0.067)	(0.059)	
Reading Index	0.075	-0.019	0.225	0.005	0.010	0.907	0.002	0.019	0.775	0.019	0.005	0.824	-0.003	0.030	0.588
	(0.077)	(0.048)		(0.053)	(0.052)		(0.059)	(0.047)		(0.058)	(0.050)		(0.063)	(0.047)	
Panel B: Dep. Vars. for Socio-Emotional Learning															
SEL Index	0.225**	0.058	0.064	0.122**	0.067	0.346	0.101*	0.096*	0.926	0.060	0.110**	0.422	0.095*	0.082*	0.824
	(0.088)	(0.049)		(0.051)	(0.056)		(0.058)	(0.049)		(0.056)	(0.052)		(0.058)	(0.050)	
Confidence score (z)	0.100	0.032	0.420	0.062	0.029	0.584	-0.001	0.079*	0.152	0.041	0.042	0.997	-0.003	0.064	0.270
	(0.079)	(0.046)		(0.049)	(0.053)		(0.051)	(0.046)		(0.054)	(0.049)		(0.053)	(0.048)	
Curiosity score (z)	0.278***	0.071	0.022	0.150***	0.088	0.271	0.177***	0.087*	0.146	0.065	0.150***	0.161	0.166***	0.084*	0.165
	(0.091)$^{+}$	(0.047)		(0.052)$^{+}$	(0.054)		(0.060)$^{+}$	(0.049)		(0.059)	(0.049)$^{+}$		(0.057)$^{+}$	(0.050)	
Panel C: Dep. Vars. for Gender Attitudes															
Gender Attitude Index	-0.112	-0.038	0.375	-0.076	-0.032	0.486	-0.111*	-0.022	0.170	0.059	-0.095*	0.007	-0.063	-0.061	0.975
	(0.088)	(0.048)		(0.057)	(0.056)		(0.058)	(0.055)		(0.065)	(0.048)		(0.052)	(0.056)	
Gender stereotype (z)	0.011	0.061	0.560	0.024	0.051	0.636	-0.048	0.099**	0.034	0.149**	-0.008	0.013	0.014	0.047	0.597
	(0.085)	(0.045)		(0.049)	(0.053)		(0.059)	(0.050)		(0.058)$^{+}$	(0.048)		(0.052)	(0.052)	
Gender roles (z)	-0.203**	-0.092**	0.162	-0.151***	-0.060	0.212	-0.149**	-0.097*	0.431	-0.055	-0.118***	0.289	-0.109**	-0.120**	0.853
	(0.082)$^{+}$	(0.045)		(0.055)$^{+}$	(0.059)		(0.058)$^{+}$	(0.053)		(0.063)	(0.045)$^{+}$		(0.052)	(0.053)	
Gender bias (z)	-0.321**	-0.054	0.093	-0.180*	-0.038	0.253	-0.246**	-0.032	0.097	0.121	-0.204**	0.005	-0.129	-0.125	0.971
	(0.161)	(0.090)		(0.107)	(0.106)		(0.116)	(0.103)		(0.123)	(0.092)		(0.098)	(0.108)	
Share	0.20	0.80		0.50	0.50		0.41	0.59		0.36	0.64		0.45	0.55	

Notes: Each coefficient corresponds to the treatment indicator from an OLS regression of equation 1, restricting the sample to the sub-group indicated by the column. All regressions include baseline dependent variable, quadruple fixed effects and school controls. School-level controls (averages): share of girls, asset score, caregiver education, grade, reading score, HH size, TV watching time, and school size. Columns indicating *p-value diff.* corresponds to testing the null hypothesis that coefficients are equal between subgroup or not. All heterogeneity characteristics are defined at baseline. Caregiver education defined as high if caregiver completed at least a high school education. Household size divided by median size. Cluster-robust standard errors at the school level in parenthesis. Significance: *** p<0.01, ** p<0.05, * p<0.1. Q-value significance (see Appendix Table A6 for values and details): + q<0.1, ++ q<0.05.

Table A5: ITT Effects - Heterogeneity *(continued)*

Heterogeneity Var.:	(16)	(17) SES	(18)	(19)	(20) Location watch TV	(21)	(22)	(23) Watch w sibling/friend	(24)	(25)	(26) Never miss school	(27)	(28)	(29) School Urban	(30)
	High	Low	p-value diff.	Home	Other	p-value diff.	Yes	No	p-value diff.	Yes	No	p-value diff.	Yes	No	p-value diff.
Panel A: Dep. Vars. for Literacy															
Literacy Index	0.039	0.074	0.555	0.066	-0.004	0.454	0.036	0.147**	0.106	0.070	-0.000	0.266	0.137	0.085	0.657
	(0.068)	(0.053)		(0.051)	(0.098)		(0.055)	(0.069)		(0.053)	(0.074)		(0.107)	(0.057)	
Comprehension Index	0.135*	0.090	0.493	0.098*	0.136	0.698	0.082	0.155**	0.281	0.090	0.064	0.693	0.052	0.119*	0.634
	(0.072)	(0.057)		(0.056)	(0.098)		(0.056)	(0.075)		(0.057)	(0.072)		(0.130)	(0.062)	
Reading Index	-0.038	0.044	0.147	0.029	-0.141	0.078	-0.006	0.095	0.127	0.044	-0.062	0.105	0.209***	0.040	0.032
	(0.062)	(0.048)		(0.044)	(0.100)		(0.050)	(0.064)		(0.044)	(0.075)		(0.065)$^{+}$	(0.047)	
Panel B: Dep. Vars. for Socio-Emotional Learning															
SEL Index	0.158**	0.064	0.171	0.102**	0.034	0.515	0.080*	0.104	0.736	0.116**	0.016	0.092	-0.039	0.122**	0.245
	(0.069)	(0.048)		(0.045)	(0.111)		(0.047)	(0.073)		(0.050)	(0.058)		(0.132)	(0.052)	
Confidence score (z)	0.091	0.021	0.324	0.060	-0.068	0.210	0.040	-0.018	0.426	0.062	-0.029	0.168	0.067	0.047	0.886
	(0.068)	(0.045)		(0.042)	(0.109)		(0.043)	(0.075)		(0.047)	(0.060)		(0.129)	(0.048)	
Curiosity score (z)	0.182***	0.090*	0.141	0.117**	0.139	0.830	0.100**	0.199***	0.163	0.139***	0.059	0.179	-0.124	0.163***	0.017
	(0.064)$^{+}$	(0.049)		(0.046)$^{+}$	(0.104)		(0.047)	(0.072)$^{+}$		(0.050)$^{+}$	(0.058)		(0.112)	(0.052)$^{+}$	
Panel C: Dep. Vars. for Gender Attitudes															
Gender Attitude Index	0.020	-0.086*	0.135	-0.061	-0.108	0.648	-0.065	-0.028	0.630	-0.088*	-0.025	0.339	-0.141	-0.085*	0.686
	(0.069)	(0.050)		(0.049)	(0.102)		(0.048)	(0.082)		(0.053)	(0.061)		(0.133)	(0.050)	
Gender stereotype (z)	0.086	0.025	0.365	0.036	-0.000	0.736	0.038	0.068	0.718	0.023	0.043	0.781	0.114	0.018	0.420
	(0.061)	(0.049)		(0.044)	(0.108)		(0.045)	(0.083)		(0.049)	(0.061)		(0.113)	(0.044)	
Gender roles (z)	-0.036	-0.142***	0.124	-0.120***	-0.099	0.832	-0.117***	-0.106	0.888	-0.129***	-0.093	0.581	-0.210*	-0.127***	0.502
	(0.066)	(0.047)$^{+}$		(0.045)$^{+}$	(0.100)		(0.045)$^{+}$	(0.084)		(0.049)$^{+}$	(0.063)		(0.116)	(0.047)$^{+}$	
Gender bias (z)	-0.064	-0.144	0.578	-0.115	-0.252	0.516	-0.107	-0.162	0.732	-0.154	-0.125	0.823	-0.224	-0.159*	0.805
	(0.136)	(0.096)		(0.092)	(0.210)		(0.090)	(0.167)		(0.098)	(0.124)		(0.256)	(0.092)	
Share	0.31	0.69		0.87	0.13		0.79	0.21		0.70	0.30		0.19	0.81	

Notes: Each coefficient corresponds to the treatment indicator from an OLS regression of equation 1, restricting the sample to the sub-group indicated by the column. All regressions include baseline dependent variable, quadruple fixed effects and school controls. School-level controls (averages): share of girls, asset score, caregiver education, grade, reading score, HH size, TV watching time, and school size. Columns indicating *p-value diff.* corresponds to testing the null hypothesis that coefficients are equal between subgroup or not. All heterogeneity characteristics are defined at baseline. SES defined by the median value of baseline household asset score. Never miss school equals one if child reported Never or Rarely missing school at baseline. Cluster-robust standard errors at the school level in parenthesis. Significance: *** p<0.01, ** p<0.05, * p<0.1. Q-value significance (see Appendix Table A6 for values and details): + q<0.1, ++ q<0.05.

Table A5: ITT Effects - Heterogeneity *(continued)*

Heterogeneity Var.:	(31)	(32)	(33)	(34)	(35)	(36)	(37)	(38)	(39)
	Baseline Outcome			Grade					
	High	Low	p-value diff.	Grade 1	Grade 2	Grade 3	p-value G1=G3	p-value G2=G3	p-value G1=G2
Panel A: Dep. Vars. for Literacy									
Literacy Index	0.093	0.022	0.272	0.079	0.036	0.028	0.279	0.678	0.516
	(0.057)	(0.062)		(0.064)	(0.066)	(0.059)			
Comprehension Index	0.086	0.097	0.875	0.103	0.087	0.067	0.547	0.707	0.832
	(0.066)	(0.064)		(0.077)	(0.068)	(0.057)			
Reading Index	0.053	-0.045	0.097	0.025	-0.015	-0.014	0.350	0.742	0.481
	(0.054)	(0.053)		(0.056)	(0.059)	(0.059)			
Panel B: Dep. Vars. for Socio-Emotional Learning									
SEL Index	0.062	0.120**	0.314	-0.012	0.121**	0.180***	0.012	0.391	0.080
	(0.052)	(0.053)		(0.064)	(0.061)	(0.057)			
Confidence score (z)	0.036	0.071	0.600	-0.051	0.057	0.105*	0.041	0.420	0.179
	(0.061)	(0.047)		(0.064)	(0.059)	(0.056)			
Curiosity score (z)	0.078	0.134***	0.370	0.032	0.151**	0.203***	0.034	0.575	0.113
	(0.063)	$(0.049)^{+}$		(0.062)	$(0.061)^{+}$	$(0.059)^{+}$			
Panel C: Dep. Vars. for Gender Attitudes									
Gender Attitude Index	-0.085	-0.042	0.517	-0.152**	0.023	-0.044	0.159	0.301	0.032
	(0.065)	(0.047)		(0.071)	(0.060)	(0.058)			
Gender stereotype (z)	0.071	0.020	0.396	0.059	0.036	0.020	0.570	0.797	0.778
	(0.058)	(0.047)		(0.067)	(0.060)	(0.057)			
Gender roles (z)	-0.103	-0.123***	0.768	-0.247***	-0.023	-0.082	0.036	0.326	0.004
	(0.065)	$(0.042)^{+}$		$(0.067)^{++}$	(0.055)	(0.059)			
Gender bias (z)	-0.059	-0.129	0.630	-0.287**	0.025	-0.099	0.272	0.284	0.047
	(0.121)	(0.106)		(0.137)	(0.113)	(0.116)			

Notes: Each coefficient corresponds to the treatment indicator from an OLS regression of equation 1, restricting the sample to the sub-group indicated by the column. All regressions include baseline dependent variable, quadruple fixed effects and school controls. School-level controls (averages): share of girls, asset score, caregiver education, grade, reading score, HH size, TV watching time, and school size. Columns indicating *p-value diff.* corresponds to testing the null hypothesis that coefficients are equal between subgroup or not. All heterogeneity characteristics are defined at baseline. Baseline outcome defined by median value of baseline outcome. Cluster-robust standard errors at the school level in parenthesis. Significance: *** p<0.01, ** p<0.05, * p<0.1. Q-value significance (see Appendix Table A6 for values and details): + q<0.1, ++ q<0.05.

Table A6: Q-values for Appendix Table A5

		(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
		Outcomes									
	Het. Var.	Literacy			SEL			Gender			
	Value	Index	Compr.	Reading	Index	Confidence	Curiosity	Index	Stereotype	Roles	Bias
Heterogeneity Variable:											
English in HH	0	0.880	0.552	0.783	0.413	0.658	0.367	0.596	0.417	0.204	0.711
	1	0.220	0.097	0.544	0.220	0.438	0.062	0.413	0.941	0.095	0.229
Boy	0	0.587	0.449	0.891	0.413	0.742	0.360	0.709	0.548	0.544	0.809
	1	0.480	0.258	0.956	0.220	0.438	0.062	0.413	0.756	0.070	0.331
HH has books	0	0.413	0.360	0.784	0.262	0.325	0.301	0.807	0.233	0.270	0.828
	1	0.709	0.367	0.984	0.324	0.984	0.062	0.270	0.581	0.080	0.195
High caregiver Education	0	0.534	0.363	0.956	0.220	0.576	0.062	0.262	0.924	0.080	0.161
	1	0.480	0.363	0.828	0.464	0.610	0.502	0.534	0.080	0.574	0.544
Large HH	0	0.434	0.354	0.697	0.324	0.417	0.331	0.464	0.569	0.151	0.457
	1	0.696	0.439	0.975	0.324	0.975	0.062	0.413	0.855	0.195	0.417
High SES	0	0.413	0.360	0.552	0.413	0.756	0.270	0.324	0.752	0.062	0.367
	1	0.709	0.270	0.704	0.220	0.417	0.070	0.828	0.398	0.740	0.756
Watch TV at home	0	0.977	0.413	0.398	0.828	0.704	0.417	0.468	1.000	0.544	0.449
	1	0.413	0.301	0.692	0.220	0.393	0.089	0.413	0.581	0.077	0.438
Watch TV with sib/fri	0	0.220	0.204	0.372	0.413	0.875	0.070	0.819	0.581	0.438	0.544
	1	0.689	0.380	0.941	0.324	0.552	0.195	0.413	0.576	0.080	0.449
Never miss school	0	0.996	0.569	0.581	0.828	0.756	0.544	0.807	0.656	0.372	0.544
	1	0.413	0.360	0.544	0.220	0.417	0.070	0.324	0.756	0.077	0.360
Urban	0	0.412	0.255	0.574	0.220	0.544	0.062	0.324	0.783	0.074	0.322
	1	0.413	0.783	0.062	0.828	0.752	0.495	0.468	0.544	0.300	0.574
High baseline outcome	0	0.819	0.367	0.574	0.220	0.367	0.070	0.535	0.780	0.062	0.439
	1	0.331	0.420	0.544	0.413	0.711	0.439	0.413	0.439	0.360	0.756
Grade	1	0.413	0.417	0.765	0.880	0.587	0.751	0.220	0.569	0.046	0.195
	2	0.722	0.438	0.867	0.262	0.548	0.095	0.811	0.711	0.783	0.883
	3	0.763	0.449	0.875	0.133	0.268	0.058	0.618	0.817	0.413	0.574

Notes: This table presents Q-values for the estimates of Appendix Table A5. Q-values correspond to False Discovery Rate (FDR)-adjusted significance levels across all tests within aggregated indices or all tests within 7 primary outcomes, following Benjamini et al. (2006).

Table A7: Effects of Encouragement on Compliance by Speaking English in Household

	(1)	(2)	(3)	(4)	(5)	
		Self-Report:			Recall:	
Dep. Var.:	Ever watch	Consistent watch	% Episode watch	Char. ID	Consistent watch (ID'd)	% Episode ID'd
Treatment	0.470***	0.201***	0.197***	0.347***	0.089***	0.100***
	(0.020)	(0.016)	(0.011)	(0.021)	(0.014)	(0.010)
Treatment X English	-0.063	0.058*	0.024	0.004	-0.022	-0.021
	(0.039)	(0.032)	(0.024)	(0.038)	(0.025)	(0.018)
Mean Control-Other	0.29	0.05	0.08	0.10	0.03	0.04
Mean Control-English	0.40	0.10	0.13	0.14	0.07	0.07
School Controls	Yes	Yes	Yes	Yes	Yes	Yes
R^2	0.318	0.168	0.239	0.258	0.119	0.177
Observations	3,662	3,662	3,662	3,662	3,662	3,662

Notes: Each column presents estimates corresponding to an OLS regression for the treatment indicator and its interaction with an indicator variable that equals one if the child's household predominantly speaks English. All estimates include quadruplet fixed effects and school-level controls – all interacted with indicator for speaking English at home. School-level controls (averages): share of girls, asset score, caregiver education, grade, reading score, HH size, TV watching time, and school size. Cluster-robust standard errors at the school level in parenthesis. Significance: *** p<0.01, ** p<0.05, * p<0.1

Table A8: Baseline Balance - Speaks English at Home Only

	(1) Treatment	(2) Control	(3) Difference	(4) p-value
Demographics & HH Characteristics				
Child is female	0.487	0.524	-0.032	0.290
Child in Grade 1	0.276	0.268	0.021	0.443
Child in Grade 2	0.357	0.333	0.025	0.488
Child in Grade 3	0.367	0.399	-0.046	0.154
Hours watch TV per day	1.366	1.424	-0.032	0.532
Watch 2+ hours per day	0.383	0.404	-0.018	0.626
Watch Akili Kids Channel	0.441	0.484	-0.042	0.409
HH size	5.435	5.459	-0.051	0.729
HH has TV	0.990	0.958	0.029	0.009
HH has smartphone	0.654	0.663	0.012	0.776
HH Asset score	4.418	4.469	-0.062	0.630
Caregiver has primary or less	0.505	0.486	0.030	0.421
HH speaks english	1.000	1.000	0.000	.
Nr books	15.217	16.052	-0.395	0.716
Main Outcomes				
Literacy Index	0.211	0.236	-0.056	0.586
Comprehension Index	0.185	0.276	-0.150	0.158
Reading Index	0.244	0.196	0.043	0.653
SEL Index	0.146	0.198	-0.075	0.217
Confidence score (z)	0.169	0.181	-0.011	0.867
Curiosity score (z)	0.103	0.179	-0.117	0.059
Gender Attitude Index	0.049	-0.009	-0.000	0.999
Gender stereotype (z)	-0.124	-0.042	-0.075	0.308
Gender roles (z)	0.095	0.001	0.030	0.693
Gender bias (z)	0.152	0.003	0.088	0.375
Secondary Outcomes				
Reading comprehension (z)	0.248	0.290	-0.093	0.309
Listening comprehension (z)	0.080	0.191	-0.160	0.126
Non-word reading (z)	0.292	0.256	0.031	0.742
Letter recognition (z)	0.194	0.134	0.082	0.377
Oral fluency (z)	0.202	0.191	-0.023	0.820
Girl Gender roles (z)	0.043	0.043	0.008	0.902
Boy Gender roles (z)	0.057	-0.015	-0.005	0.943
Boy Gender bias: willing to interact (z)	-0.165	-0.221	0.057	0.482
Girl Gender bias: willing to interact (z)	-0.245	-0.392	0.120	0.161
Boy Gender bias: interactions (z)	-0.145	-0.191	0.027	0.666
Girl Gender bias: interactions (z)	-0.194	-0.327	0.111	0.144

Notes: Variable means for the treatment and control groups in columns (1) and (2). Columns (3) and (4) present the estimates for the adjusted differences in means and corresponding p-value for a significance test. The difference was calculated from an OLS regression of the treatment indicator on the balance variable, adjusting for quadruplet fixed effects. Standard errors are clustered at the school level.

Table A9: Selecting Variables for Heterogeneity with Causal Forests - Variable Importance

Dep. Var.:	(1) Literacy Index	(2) SEL Index	(3) Gender Index
English in HH	0.046	0.035	0.016
Boy	0.008	0.017	0.024
HH has books	0.012	0.015	0.022
Caregiver has Higher Ed	0.007	0.013	0.027
Large HH	0.011	0.016	0.013
High SES	0.009	0.016	0.033
Watch TV at home	0.020	0.009	0.009
Watch w/Sibling or friend	0.005	0.006	0.009
Never missed school	0.009	0.027	0.018
Urban School	0.101	0.041	0.099
Grade	0.018	0.039	0.039
Quadruplet	0.610	0.599	0.449
Baseline outcome	0.142	0.164	0.238
Outcome high	0.004	0.004	0.004

Notes: Variable importance from a school-level clustered causal forest following Wager and Athey (2018) for each outcome. All independent variables are baseline values.

Table A10: Heterogeneity Estimates using Best Linear Projections (BLP)

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
	Literacy Index			SEL Index			Gender Index		
	Estimate	Std. Error	p-value	Estimate	Std. Error	p-value	Estimate	Std. Error	p-value
English in HH	0.224*	(0.119)	0.0597	0.214**	(0.098)	0.0293	-0.052	(0.111)	0.6404
Boy	-0.017	(0.069)	0.8094	0.049	(0.069)	0.4772	-0.066	(0.080)	0.4092
HH has books	0.081	(0.081)	0.3190	-0.022	(0.079)	0.7802	-0.084	(0.082)	0.3061
Caregiver has Higher Ed	-0.039	(0.069)	0.5693	-0.064	(0.077)	0.4009	0.138*	(0.070)	0.0500
Large HH	-0.032	(0.088)	0.7189	0.046	(0.071)	0.5129	-0.014	(0.073)	0.8451
High SES	-0.044	(0.075)	0.5612	0.058	(0.084)	0.4936	0.149	(0.087)	0.0883
Watch TV at home	0.110	(0.210)	0.6009	0.004	(0.124)	0.9758	-0.021	(0.118)	0.8608
Watch w/Sibling or friend	-0.108	(0.101)	0.2850	-0.039	(0.085)	0.6463	-0.012	(0.092)	0.8967
Never missed school	0.027	(0.079)	0.7316	0.155*	(0.076)	0.0419	-0.066	(0.086)	0.4389
Urban School	-0.253	(0.258)	0.3264	-0.132	(0.140)	0.3453	0.240	(0.149)	0.1087
Grade	-0.086*	(0.049)	0.0773	0.046	(0.047)	0.3336	0.046	(0.045)	0.3002
Baseline outcome	0.006	(0.063)	0.9215	-0.096	(0.064)	0.1322	0.127**	(0.056)	0.0245
Outcome high	0.123	(0.110)	0.2638	0.076	(0.116)	0.5117	-0.202*	(0.110)	0.0670
Intercept	0.137	(0.290)	0.6368	-0.165	(0.236)	0.4860	-0.021	(0.232)	0.9272

Notes: Estimates for the best linear projection (BLP) of a conditional average treatment effect following Semenova and Chernozhukov (2021). Estimates based on Wager and Athey (2018) using a school-level clustered causal forest. All independent variables are baseline values. Significance: *** p<0.01, ** p<0.05, * p<0.1

Table A11: Effects of Encouragement on Compliance by Having a TV at Home

	(1)	(2)	(3)	(4)	(5)	
		Self-Report:			Recall:	
Dep. Var.:	Ever watch	Consistent watch	% Episode watch	Char. ID	Consistent watch (ID'd)	% Episode ID'd
Treatment	0.479***	0.209***	0.205***	0.319***	0.083***	0.104***
	(0.042)	(0.034)	(0.021)	(0.038)	(0.023)	(0.015)
Treatment X Home TV	-0.021	0.005	-0.002	0.041	0.005	-0.004
	(0.044)	(0.036)	(0.022)	(0.040)	(0.026)	(0.017)
Mean Control-No Home TV	0.29	0.05	0.08	0.10	0.03	0.04
Mean Control-Home TV	0.40	0.10	0.13	0.15	0.07	0.08
School Controls	Yes	Yes	Yes	Yes	Yes	Yes
R^2	0.311	0.153	0.226	0.238	0.106	0.171
Observations	3,654	3,654	3,654	3,654	3,654	3,654

Notes: Each column presents estimates corresponding to an OLS regression for the treatment indicator and its interaction with an indicator variable that equals one if the child has a TV in the household. All estimates include quadruplet fixed effects and school-level controls – all interacted with indicator for having a TV in the household. School-level controls (averages): share of girls, asset score, caregiver education, grade, reading score, HH size, TV watching time, and school size. Cluster-robust standard errors at the school level in parenthesis. Significance: *** p<0.01, ** p<0.05, * p<0.1

Table A12: Heterogeneity by Intensity of Energy Outages - Compliance Measures

	(1)	(2)	(3)	(4)	(5)	
		Self-Report:			Recall:	
Dep. Var.:	Ever watch	Consistent watch	% Episode watch	Char. ID	Consistent watch (ID'd)	% Episode ID'd
Treatment	0.498***	0.210***	0.217***	0.410***	0.084***	0.108***
	(0.023)	(0.024)	(0.015)	(0.027)	(0.019)	(0.013)
Treat. X Outage	-0.114***	-0.030	-0.054**	-0.135***	0.004	-0.019
	(0.040)	(0.032)	(0.023)	(0.040)	(0.031)	(0.023)
Mean Control	0.31	0.06	0.09	0.11	0.04	0.05
School Controls	Yes	Yes	Yes	Yes	Yes	Yes
R^2	0.333	0.181	0.250	0.270	0.135	0.211
Observations	3,668	3,668	3,668	3,668	3,668	3,668

Notes: Each column presents estimates corresponding to an OLS regression for the treatment indicator and its interaction with an indicator variable that equals one if the child is living in a county identified as having a high amount of electricity outages during the study period. To classify counties, we use the average SAIDI (System Average Interruption Duration Index) that captures the total duration of electricity outages experienced by the average customer. All estimates include the un-interacted indicator for high outage, quadruplet fixed effects and school-level controls. School-level controls (averages): share of girls, asset score, caregiver education, grade, reading score, HH size, TV watching time, and school size. Cluster-robust standard errors at the school level in parenthesis. Significance: *** p<0.01, ** p<0.05, * p<0.1

Table A13: Heterogeneity by Intensity of Energy Outages - Primary Outcomes

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
Dep. Var.:	Literacy Index	Comprehension	Reading	SEL Index	Confidence	Curiosity	Gender Index	Stereotype	Roles	Bias
Treatment	0.143*	0.218***	0.028	0.041	-0.018	0.095	-0.155**	-0.078	-0.188***	-0.322**
	(0.077)	(0.071)	(0.076)	(0.068)	(0.067)	(0.062)	(0.072)	(0.070)	(0.069)	(0.140)
Treatment X Outage	-0.029	-0.079	0.018	0.108	0.115	0.071	0.111	0.187**	0.063	0.280
	(0.104)	(0.114)	(0.095)	(0.093)	(0.086)	(0.089)	(0.101)	(0.093)	(0.097)	(0.188)
q-value Treat.	0.191	0.033	0.831	0.658	0.842	0.320	0.188	0.460	0.050	0.104
q-value Treat. X Out.	0.777	0.653	0.850	0.408	0.371	0.653	0.408	0.155	0.653	0.320
School Controls	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
R^2	0.437	0.371	0.440	0.138	0.121	0.125	0.228	0.132	0.172	0.247
Observations	3,668	3,668	3,668	3,668	3,668	3,668	3,668	3,668	3,668	3,668

Notes: Each column presents estimates corresponding to an OLS regression for the treatment indicator and its interaction with an indicator variable that equals one if the child is living in a county identified as having a high amount of electricity outages during the study period. To classify counties, we use the average SAIDI (System Average Interruption Duration Index) that captures the total duration of electricity outages experienced by the average customer. All estimates include the un-interacted indicator for high outage, quadruplet fixed effects and school-level controls. School-level controls (averages): share of girls, asset score, caregiver education, grade, reading score, HH size, TV watching time, and school size. Q-values in column (4) correspond to False Discovery Rate (FDR)-adjusted significance levels within each family of outcomes, following Benjamini et al. (2006). Cluster-robust standard errors at the school level in parenthesis. Significance: *** p<0.01, ** p<0.05, * p<0.1

Table A14: Heterogeneity in Compliance

	(1)	(2)	(3)	(4)	(5)	(6)
	Self-Report			Recall-Based		
Dep Var.:	Ever Watch	Consistent Watch	% Episodes Watch	Character Identification	Consistent Watch (ID'd)	% Episodes Identified
English at home (=1)	0.407***	0.259***	0.221***	0.351***	0.067**	0.079***
	(0.039)$^{+++}$	(0.030)$^{+++}$	(0.023)$^{+++}$	(0.039)$^{+++}$	(0.027)$^{++}$	(0.020)$^{+++}$
English at home (=0)	0.470***	0.201***	0.197***	0.347***	0.089***	0.100***
	(0.020)$^{+++}$	(0.016)$^{+++}$	(0.011)$^{+++}$	(0.021)$^{+++}$	(0.013)$^{+++}$	(0.010)$^{+++}$
diff. p-value	0.105	0.073	0.318	0.916	0.367	0.235
Boy (=1)	0.455***	0.218***	0.202***	0.374***	0.080***	0.093***
	(0.023)$^{+++}$	(0.019)$^{+++}$	(0.013)$^{+++}$	(0.023)$^{+++}$	(0.017)$^{+++}$	(0.011)$^{+++}$
Boy (=0)	0.463***	0.203***	0.202***	0.337***	0.093***	0.102***
	(0.023)$^{+++}$	(0.018)$^{+++}$	(0.012)$^{+++}$	(0.023)$^{+++}$	(0.014)$^{+++}$	(0.011)$^{+++}$
diff. p-value	0.776	0.482	0.990	0.111	0.454	0.361
HH has books (=1)	0.472***	0.222***	0.219***	0.376***	0.078***	0.101***
	(0.024)$^{+++}$	(0.020)$^{+++}$	(0.015)$^{+++}$	(0.023)$^{+++}$	(0.015)$^{+++}$	(0.012)$^{+++}$
HH has books (=0)	0.451***	0.212***	0.196***	0.348***	0.092***	0.099***
	(0.023)$^{+++}$	(0.018)$^{+++}$	(0.013)$^{+++}$	(0.023)$^{+++}$	(0.016)$^{+++}$	(0.011)$^{+++}$
diff. p-value	0.444	0.663	0.172	0.293	0.408	0.856
Caregiver high educ. (=1)	0.450***	0.203***	0.200***	0.392***	0.105***	0.110***
	(0.026)$^{+++}$	(0.021)$^{+++}$	(0.014)$^{+++}$	(0.027)$^{+++}$	(0.021)$^{+++}$	(0.015)$^{+++}$
Caregiver high educ. (=0)	0.467***	0.214***	0.203***	0.338***	0.080***	0.094***
	(0.022)$^{+++}$	(0.017)$^{+++}$	(0.012)$^{+++}$	(0.020)$^{+++}$	(0.013)$^{+++}$	(0.010)$^{+++}$
diff. p-value	0.561	0.652	0.867	0.041	0.216	0.218
Large HH (=1)	0.458***	0.212***	0.206***	0.356***	0.083***	0.100***
	(0.025)$^{+++}$	(0.019)$^{+++}$	(0.013)$^{+++}$	(0.022)$^{+++}$	(0.015)$^{+++}$	(0.011)$^{+++}$
Large HH (=0)	0.459***	0.208***	0.196***	0.353***	0.084***	0.095***
	(0.024)$^{+++}$	(0.018)$^{+++}$	(0.012)$^{+++}$	(0.024)$^{+++}$	(0.016)$^{+++}$	(0.012)$^{+++}$
diff. p-value	0.956	0.858	0.510	0.917	0.968	0.730
High SES (=1)	0.462***	0.215***	0.206***	0.403***	0.099***	0.106***
	(0.026)$^{+++}$	(0.024)$^{+++}$	(0.016)$^{+++}$	(0.028)$^{+++}$	(0.019)$^{+++}$	(0.014)$^{+++}$
High SES (=0)	0.462***	0.214***	0.201***	0.338***	0.081***	0.095***
	(0.022)$^{+++}$	(0.018)$^{+++}$	(0.012)$^{+++}$	(0.021)$^{+++}$	(0.014)$^{+++}$	(0.010)$^{+++}$
diff. p-value	0.991	0.958	0.787	0.027	0.359	0.415
Watch TV at home (=1)	0.457***	0.214***	0.203***	0.360***	0.089***	0.100***
	(0.020)$^{+++}$	(0.015)$^{+++}$	(0.011)$^{+++}$	(0.020)$^{+++}$	(0.014)$^{+++}$	(0.011)$^{+++}$
Watch TV at home (=0)	0.479***	0.209***	0.205***	0.319***	0.083***	0.104***
	(0.045)$^{+++}$	(0.037)$^{+++}$	(0.022)$^{+++}$	(0.040)$^{+++}$	(0.025)$^{+++}$	(0.016)$^{+++}$
diff. p-value	0.631	0.890	0.927	0.308	0.835	0.799

Notes: Each coefficient corresponds to the treatment indicator from an OLS regression of equation 1, restricting the sample to the sub-group indicated by the row. Columns indicate corresponding dependent variable. All regressions include quadruple fixed effects and school controls. School-level controls (averages): share of girls, asset score, caregiver education, grade, reading score, HH size, TV watching time, and school size. P-values for testing the null hypothesis that coefficients are equal between subgroup or not presented in rows labeled *diff. p-value*. All heterogeneity characteristics are defined at baseline. Caregiver high education indicates whether caregiver has completed high school or higher. SES defined by the median value of baseline household asset score. Never miss school equals one if child reported Never or Rarely missing school at baseline. Cluster-robust standard errors at the school level in parenthesis. Significance: *** p<0.01, ** p<0.05, * p<0.1. Q-value significance: + q<0.1, ++ q<0.05.

Table A14: Heterogeneity in Compliance *(continued)*

	(1)	(2)	(3)	(4)	(5)	(6)
		Self-Report			Recall-Based	
Dep Var.:	Ever Watch	Consistent Watch	% Episodes Watch	Character Identification	Consistent Watch (ID'd)	% Episodes Identified
Watch TV w/sib./fr. (=1)	0.477***	0.222***	0.213***	0.348***	0.085***	0.097***
	$(0.019)^{+++}$	$(0.017)^{+++}$	$(0.011)^{+++}$	$(0.021)^{+++}$	$(0.014)^{+++}$	$(0.010)^{+++}$
Watch TV w/sib./fr. (=0)	0.403***	0.182***	0.168***	0.398***	0.083***	0.098***
	$(0.039)^{+++}$	$(0.025)^{+++}$	$(0.018)^{+++}$	$(0.035)^{+++}$	$(0.022)^{+++}$	$(0.015)^{+++}$
diff. p-value	0.046	0.146	0.020	0.156	0.933	0.938
Never miss school (=1)	0.458***	0.209***	0.200***	0.345***	0.081***	0.098***
	$(0.020)^{+++}$	$(0.016)^{+++}$	$(0.011)^{+++}$	$(0.022)^{+++}$	$(0.013)^{+++}$	$(0.010)^{+++}$
Never miss school (=0)	0.457***	0.207***	0.198***	0.382***	0.104***	0.099***
	$(0.032)^{+++}$	$(0.025)^{+++}$	$(0.017)^{+++}$	$(0.027)^{+++}$	$(0.021)^{+++}$	$(0.016)^{+++}$
diff. p-value	0.972	0.925	0.902	0.191	0.218	0.901
Urban (=1)	0.432***	0.226***	0.199***	0.252***	0.127***	0.116***
	$(0.073)^{+++}$	$(0.041)^{+++}$	$(0.031)^{+++}$	$(0.064)^{+++}$	$(0.046)^{+++}$	$(0.031)^{+++}$
Urban (=0)	0.488***	0.228***	0.220***	0.391***	0.092***	0.107***
	$(0.020)^{+++}$	$(0.017)^{+++}$	$(0.012)^{+++}$	$(0.021)^{+++}$	$(0.015)^{+++}$	$(0.011)^{+++}$
diff. p-value	0.444	0.964	0.503	0.034	0.466	0.777
High Literacy (=1)	0.477***	0.258***	0.226***	0.387***	0.115***	0.119***
	$(0.023)^{+++}$	$(0.020)^{+++}$	$(0.014)^{+++}$	$(0.025)^{+++}$	$(0.017)^{+++}$	$(0.013)^{+++}$
High Literacy (=0)	0.441***	0.173***	0.180***	0.323***	0.057***	0.077***
	$(0.024)^{+++}$	$(0.017)^{+++}$	$(0.012)^{+++}$	$(0.025)^{+++}$	$(0.015)^{+++}$	$(0.011)^{+++}$
diff. p-value	0.209	0.000	0.004	0.034	0.002	0.002
Grade 1 (=1)	0.428***	0.180***	0.174***	0.318***	0.080***	0.090***
	(0.029)	(0.023)	(0.015)	(0.030)	(0.022)	(0.015)
Grade 2 (=1)	0.471***	0.186***	0.193***	0.344***	0.052***	0.077***
	(0.026)	(0.021)	(0.015)	(0.025)	(0.017)	(0.012)
Grade 3 (=1)	0.463***	0.266***	0.231***	0.378***	0.111***	0.118***
	(0.028)	(0.021)	(0.015)	(0.029)	(0.016)	(0.011)
G1=G3 p-value	0.315	0.004	0.003	0.113	0.220	0.089
G2=G3 p-value	0.810	0.003	0.044	0.312	0.003	0.003
G1=G2 p-value	0.216	0.822	0.299	0.435	0.204	0.381

Notes: Each coefficient corresponds to the treatment indicator from an OLS regression of equation 1, restricting the sample to the sub-group indicated by the row. Columns indicate corresponding dependent variable. All regressions include quadruple fixed effects and school controls. School-level controls (averages): share of girls, asset score, caregiver education, grade, reading score, HH size, TV watching time, and school size. P-values for testing the null hypothesis that coefficients are equal between subgroup or not presented in rows labeled *diff. p-value*. All heterogeneity characteristics are defined at baseline. Watch TV w/sib./fr. indicates whether child reported watching TV with sibling or friend. High literacy defined by median literacy index at baseline. Cluster-robust standard errors at the school level in parenthesis. Significance: *** p<0.01, ** p<0.05, * p<0.1. Q-value significance: + q<0.1, ++ q<0.05.

Table A15: Selecting Variables for Heterogeneity in Compliance with Causal Forests - Variable Importance

	(1)	(2)	(3)	(4)	(5)	(6)
		Self-Reported			Recall-Based	
Dep. Var.:	Ever watch	Consistent watch	% Episodes watch	Identify characters	Consistent watch (ID)	% Episodes identified
Urban School	0.131	0.020	0.074	0.128	0.016	0.030
Reading affect index	0.128	0.138	0.090	0.099	0.102	0.087
Literacy Index	0.120	0.431	0.297	0.154	0.285	0.281
Social interaction index	0.090	0.062	0.070	0.127	0.102	0.071
TV hours	0.050	0.015	0.018	0.014	0.014	0.014
High Outage	0.049	0.015	0.023	0.060	0.053	0.035
Confidence	0.047	0.028	0.027	0.021	0.026	0.034
Watch w/Sibling or friend	0.025	0.013	0.025	0.004	0.003	0.002
Curiosity	0.024	0.014	0.014	0.024	0.018	0.015
Large HH	0.009	0.005	0.007	0.006	0.006	0.006
Grade	0.008	0.025	0.014	0.009	0.012	0.010
Caregiver has Higher Ed	0.007	0.005	0.005	0.021	0.004	0.006
English in HH	0.006	0.013	0.012	0.017	0.007	0.006
HH has books	0.006	0.009	0.013	0.008	0.005	0.003
Watch TV at home	0.006	0.001	0.001	0.006	0.000	0.001
High SES	0.005	0.010	0.010	0.012	0.019	0.014
Boy	0.005	0.006	0.006	0.006	0.009	0.008
Never missed school	0.004	0.004	0.004	0.004	0.003	0.002
Ever ask for books	0.004	0.004	0.004	0.007	0.007	0.008
Enjoy reading	0.003	0.002	0.002	0.002	0.003	0.004
Parents encourage reading	0.001	0.000	0.001	0.002	0.003	0.001
Quadruplet	0.271	0.179	0.285	0.269	0.304	0.361

Notes: Variable importance from a school-level clustered causal forest following Wager and Athey (2018) for each outcome. All independent variables are baseline values.

Table A16: Heterogeneity in Compliance Estimates using Best Linear Projections (BLP)

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)	(13)	(14)	(15)	(16)	(17)	(18)
	Self-Report									Recall-Based								
	Ever watch			Consistent watch			% Episodes watch			Identify			Consistent watch			% Episodes		
	Estimate	S.E.	p-val.	Estimate	S.E.	p-val.	Estimate	S.E.	p-val.	Estimate	S.E.	p-val.	Estimate	S.E.	p-val.	Estimate	S.E.	p-val.
English in HH	-0.022	(0.043)	0.612	0.043	(0.039)	0.272	0.022	(0.027)	0.409	0.060	(0.045)	0.180	-0.027	(0.029)	0.344	-0.011	(0.020)	0.601
Boy	-0.016	(0.028)	0.565	0.020	(0.025)	0.437	0.001	(0.017)	0.944	0.038	(0.027)	0.156	-0.014	(0.019)	0.450	-0.011	(0.013)	0.413
HH has books	0.017	(0.034)	0.615	0.000	(0.030)	0.989	0.015	(0.021)	0.473	-0.001	(0.034)	0.973	-0.016	(0.020)	0.422	-0.003	(0.013)	0.844
Caregiver has Higher Ed	0.002	(0.036)	0.946	-0.009	(0.029)	0.754	0.003	(0.020)	0.882	0.058*	(0.033)	0.076	-0.003	(0.021)	0.902	0.009	(0.015)	0.546
Large HH	-0.025	(0.036)	0.489	-0.002	(0.025)	0.947	0.002	(0.018)	0.928	0.002	(0.030)	0.952	0.005	(0.022)	0.812	0.007	(0.015)	0.644
High SES	-0.005	(0.035)	0.894	0.006	(0.033)	0.867	0.006	(0.023)	0.780	0.036	(0.034)	0.298	0.032	(0.028)	0.256	0.016	(0.019)	0.417
Watch TV at home	0.005	(0.060)	0.938	0.011	(0.038)	0.770	0.003	(0.026)	0.920	0.031	(0.051)	0.549	-0.004	(0.025)	0.872	0.001	(0.018)	0.935
Watch w/Sibling or friend	0.074*	(0.042)	0.079	0.053	(0.034)	0.115	0.048	(0.023)	0.043	-0.005	(0.039)	0.905	-0.005	(0.025)	0.851	-0.001	(0.016)	0.942
Never missed school	0.013	(0.035)	0.712	0.003	(0.029)	0.910	0.005	(0.020)	0.816	-0.016	(0.032)	0.612	-0.008	(0.023)	0.723	0.002	(0.016)	0.901
Urban School	-0.123*	(0.069)	0.075	-0.021	(0.049)	0.668	-0.051	(0.036)	0.164	-0.119**	(0.057)	0.038	-0.025	(0.038)	0.513	-0.036	(0.028)	0.210
Grade	0.000	(0.019)	0.982	0.009	(0.017)	0.607	0.007	(0.011)	0.535	0.007	(0.020)	0.735	-0.006	(0.014)	0.685	-0.003	(0.010)	0.778
High Outage	-0.054	(0.046)	0.244	-0.038	(0.036)	0.296	-0.025	(0.025)	0.315	-0.091**	(0.045)	0.045	0.034	(0.032)	0.282	0.008	(0.025)	0.742
Literacy Index	0.025	(0.020)	0.210	0.063***	(0.017)	0.000	0.034***	(0.012)	0.004	0.026	(0.020)	0.186	0.031***	(0.011)	0.005	0.021***	(0.008)	0.011
Curiosity	0.003	(0.019)	0.878	0.004	(0.015)	0.806	0.010	(0.010)	0.324	-0.034**	(0.016)	0.035	-0.006	(0.011)	0.569	-0.002	(0.008)	0.819
Confidence	-0.034*	(0.019)	0.077	-0.025	(0.016)	0.118	-0.022**	(0.011)	0.042	-0.016	(0.016)	0.314	-0.011	(0.012)	0.336	-0.014*	(0.008)	0.068
TV hours	-0.045*	(0.020)	0.024	-0.016	(0.016)	0.308	-0.016	(0.011)	0.135	0.001	(0.018)	0.947	-0.016	(0.013)	0.224	-0.006	(0.010)	0.506
Enjoy reading	-0.018	(0.045)	0.682	-0.038	(0.036)	0.296	-0.016	(0.025)	0.530	-0.013	(0.040)	0.739	-0.024	(0.026)	0.363	-0.023	(0.018)	0.211
Parents encourage reading	-0.001	(0.041)	0.978	-0.018	(0.033)	0.584	0.004	(0.024)	0.876	0.018	(0.042)	0.673	0.028	(0.030)	0.352	0.010	(0.019)	0.579
Ever ask for books	-0.004	(0.041)	0.927	0.027	(0.032)	0.396	0.008	(0.023)	0.741	0.032	(0.040)	0.423	-0.025	(0.028)	0.385	-0.025	(0.020)	0.194
Reading affect index	-0.007	(0.019)	0.728	0.006	(0.016)	0.684	0.000	(0.011)	0.977	0.013	(0.017)	0.460	0.009	(0.013)	0.469	0.012	(0.008)	0.134
Social interaction index	0.024	(0.019)	0.205	0.014	(0.014)	0.319	0.014	(0.010)	0.142	0.028*	(0.017)	0.094	0.018	(0.011)	0.114	0.012	(0.008)	0.150
Intercept	0.615***	(0.133)	0.000	0.205**	(0.091)	0.025	0.194***	(0.068)	0.005	0.280**	(0.124)	0.024	0.184**	(0.091)	0.042	0.168***	(0.064)	0.008

Notes: Estimates for the best linear projection (BLP) of a conditional average treatment effect following Semenova and Chernozhukov (2021). Estimates based on Wager and Athey (2018) using a school-level clustered causal forest. All independent variables are baseline values. Significance: *** p<0.01, ** p<0.05, * p<0.1

Table A17: Heterogeneity in Mechanisms by Speaking English in Household

	(1)	(2)	(3)
	Language at home		
	English	Other	p-value diff.
Panel A: Preferences			
Ever Read for Enjoyment (=1)	0.015	0.025	0.718
	(0.028)	(0.016)	
Ever Ask for Books (=1)	-0.028	-0.004	0.387
	(0.025)	(0.016)	
Positive Affect Index	-0.076	-0.016	0.197
	(0.046)	(0.031)	
Panel B: Time Use			
Watch 2+ hours TV a day (=1)	0.095**	0.030	0.123
	(0.047)	(0.025)	
Panel C: Indirect Channels			
Parent Encourage (=1)	0.046*	0.009	0.178
	(0.026)	(0.013)	
Parents Read Stories (=1)	0.026	-0.020	0.285
	(0.044)	(0.020)	
Social Interaction Index	-0.119**	0.030	0.006
	(0.054)	(0.027)	
Share	0.20	0.80	

Notes: Each column presents estimates corresponding to an OLS regression for the treatment indicator and its interaction with an indicator variable that equals one if the child's household predominantly speaks English. All estimates include quadruplet fixed effects and school-level controls. School-level controls (averages): share of girls, asset score, caregiver education, grade, reading score, HH size, TV watching time, and school size. Cluster-robust standard errors at the school level in parenthesis. Significance: *** p<0.01, ** p<0.05, * p<0.1

Table A18: ITT Estimates on Child Activities during Non-Show Days at Show Times

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
Dep. Var.:	Play	Study	Homework	Chores	TV	Pray	Remedial Class	Church	Other
Treatment	0.004	-0.031	-0.037	-0.033	0.057**	-0.012	0.006	-0.038	-0.004
	(0.032)	(0.020)	(0.026)	(0.024)	(0.025)	(0.014)	(0.007)	(0.024)	(0.006)
q-value	0.902	0.316	0.316	0.316	0.197	0.560	0.560	0.316	0.560
Control Mean	0.45	0.26	0.50	0.50	0.35	0.10	0.01	0.76	0.02
School Controls	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
R^2	0.190	0.136	0.133	0.171	0.300	0.489	0.076	0.061	0.088
Observations	3,669	3,669	3,669	3,669	3,669	3,669	3,669	3,669	3,669

Notes: Each column corresponds to the count of times the caregiver reported that the child engaged in that activity on Tuesday and Sunday (non-show days) during show time. Multiple selections were allowed. Each column presents estimates for an OLS regression for the treatment indicator. All estimates include quadruplet fixed effects and school-level controls. School-level controls (averages): share of girls, asset score, caregiver education, grade, reading score, HH size, TV watching time, and school size. Q-values in column (4) correspond to False Discovery Rate (FDR)-adjusted significance levels within each family of outcomes, following Benjamini et al. (2006). Cluster-robust standard errors at the school level in parenthesis. Significance: *** p<0.01, ** p<0.05, * p<0.1

Table A19: Heterogeneity in Gender Outcomes by Whether Children Related to Character

Dep. Var.:	(1) Gender Index	(2) Stereotype	(3) Roles	(4) Bias
Treatment	-0.086	0.055	-0.159***	-0.192*
	(0.058)	(0.058)	(0.054)	(0.112)
Nuzo	-0.549	0.103	-0.694	-0.479
	(0.519)	(0.564)	(0.533)	(1.022)
Namia	0.578	0.139	0.876*	1.218
	(0.432)	(0.465)	(0.473)	(0.935)
Bubelang	0.468	0.448	0.243	0.087
	(0.634)	(0.689)	(0.582)	(1.321)
Treatment × Nuzo	0.019	-0.013	-0.010	0.051
	(0.076)	(0.078)	(0.078)	(0.155)
Treatment × Namia	0.064	-0.059	0.151**	0.187
	(0.061)	(0.066)	(0.067)	(0.125)
Treatment × Bubelang	-0.080	-0.050	-0.058	-0.196
	(0.093)	(0.094)	(0.093)	(0.184)
q-value Treat	0.179	0.345	0.014	0.172
q-value Nuzo	0.580	0.855	0.580	0.852
q-value Namia	0.257	0.765	0.256	0.257
q-value Bubelang	0.901	0.901	0.901	0.948
q-value Treat × Nuzo	0.907	0.907	0.907	0.907
q-value Treat × Namia	0.375	0.375	0.092	0.268
q-value Treat × Bubelang	0.598	0.598	0.598	0.598
School Controls	Yes	Yes	Yes	Yes
R^2	0.228	0.132	0.172	0.247
Observations	3,668	3,668	3,668	3,668

Notes: Each column presents estimates corresponding to an OLS regression for the treatment indicator and its interaction with indicator variables that equal one if the child related "somewhat" or more with the respective character. All estimates include quadruplet fixed effects and school-level controls. School-level controls (averages): share of girls, asset score, caregiver education, grade, reading score, HH size, TV watching time, and school size. Q-values correspond to False Discovery Rate (FDR)-adjusted significance levels following Benjamini et al. (2006). Cluster-robust standard errors at the school level in parenthesis. Significance: *** p<0.01, ** p<0.05, * p<0.1

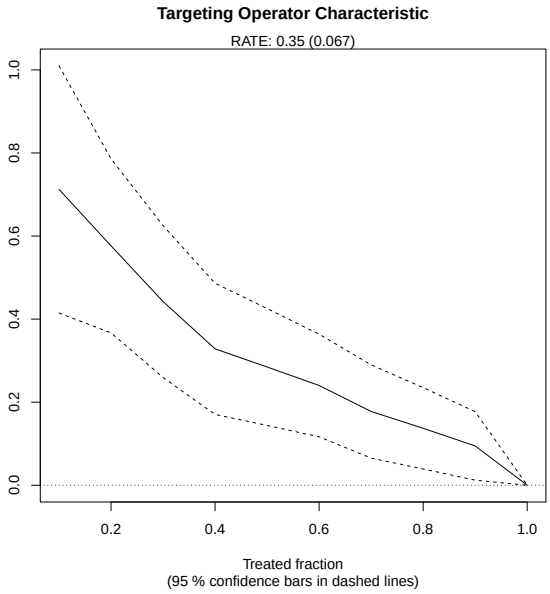
Table A20: Primary Costs Included in CEA, by Detailed Category

Cost Category	(1) Cost (2024 Real USD)	(2) Cost (2011 Real USD)	(3) Nr of entries	(4) Share of total cost	(5) Share of entries
Recurring Treatment Specific Costs					
Electricity for TV	$1,051,778.76	$768,377.54	1	36.72%	0.02%
SMS Encouragement	$804,608.49	$587,807.17	36	28.09%	0.73%
Channel Airing Costs	$71,048.79	$51,904.73	1	2.48%	0.02%
Software & Communications	$3,127.76	$2,284.99	91	0.11%	1.84%
Staff Welfare	$24,655.45	$18,012.05	1,151	0.86%	23.25%
Rent	$8,788.18	$6,420.21	99	0.31%	0.02%
Interest Expense	$3,194.94	$2,334.06	9	0.11%	0.18%
Office Expenses	$2,490.84	$1,819.68	106	0.09%	2.14%
SUM Recurring Costs	$1,969,693.21	$1,438,960.43	1,494	68.77%	28.2%
Production Costs					
Filming, Animation & Editing	$719,392.8	$525,552.8	2040	25.12%	41.21%
Admin	$53,806.56	$39,308.41	207	1.88%	4.18%
Writing	$50,886.48	$37,175.15	918	1.78%	18.55%
DSA	$32,136.2	$23,477.12	68	1.12%	1.37%
Transport	$15,873.48	$11,596.38	50	0.55%	1.01%
Background research	$10,071.46	$7,357.71	59	0.35%	1.19%
User testing	$4,829.06	$3,527.88	39	0.17%	0.79%
Accommodation	$3,615.79	$2,641.52	6	0.13%	0.12%
Meals	$1,764.27	$1,288.89	55	0.06%	1.11%
Travel	$1,368.11	$999.47	2	0.05%	0.04%
Design	$504.85	$368.82	2	0.02%	0.04%
Staff Trainings	$157.77	$115.26	9	0.01%	0.18%
Fuel expenses	$83.5	$61	1	0%	0.02%
SUM Production Costs	$894,490.33	$653,470.41	3,456	31.24%	69.81%
Annualized Production Costs	$89,449.03	$65,347.04			

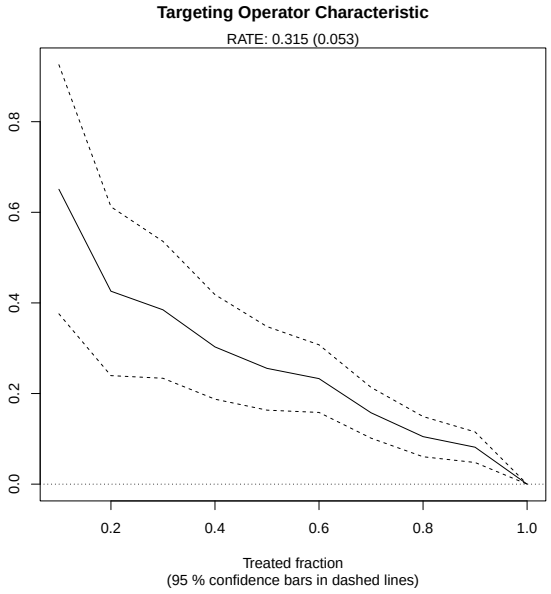
Note: For our cost-effectiveness analysis, investment costs were annualized by dividing by 10, in accordance with UK GAAP guidelines as outlined in Section 6. Production/Investment costs represent one-time expenses incurred during show production, while recurring costs would be incurred annually throughout the applicable timeframe.

Figure A1: Assessing Heterogeneity with Rank-Weighted Average Treatment Effects (RATE)

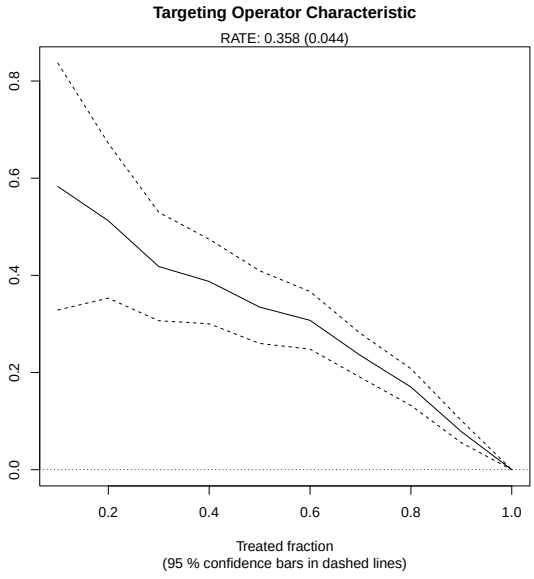
a. Literacy Index



b. SEL Index

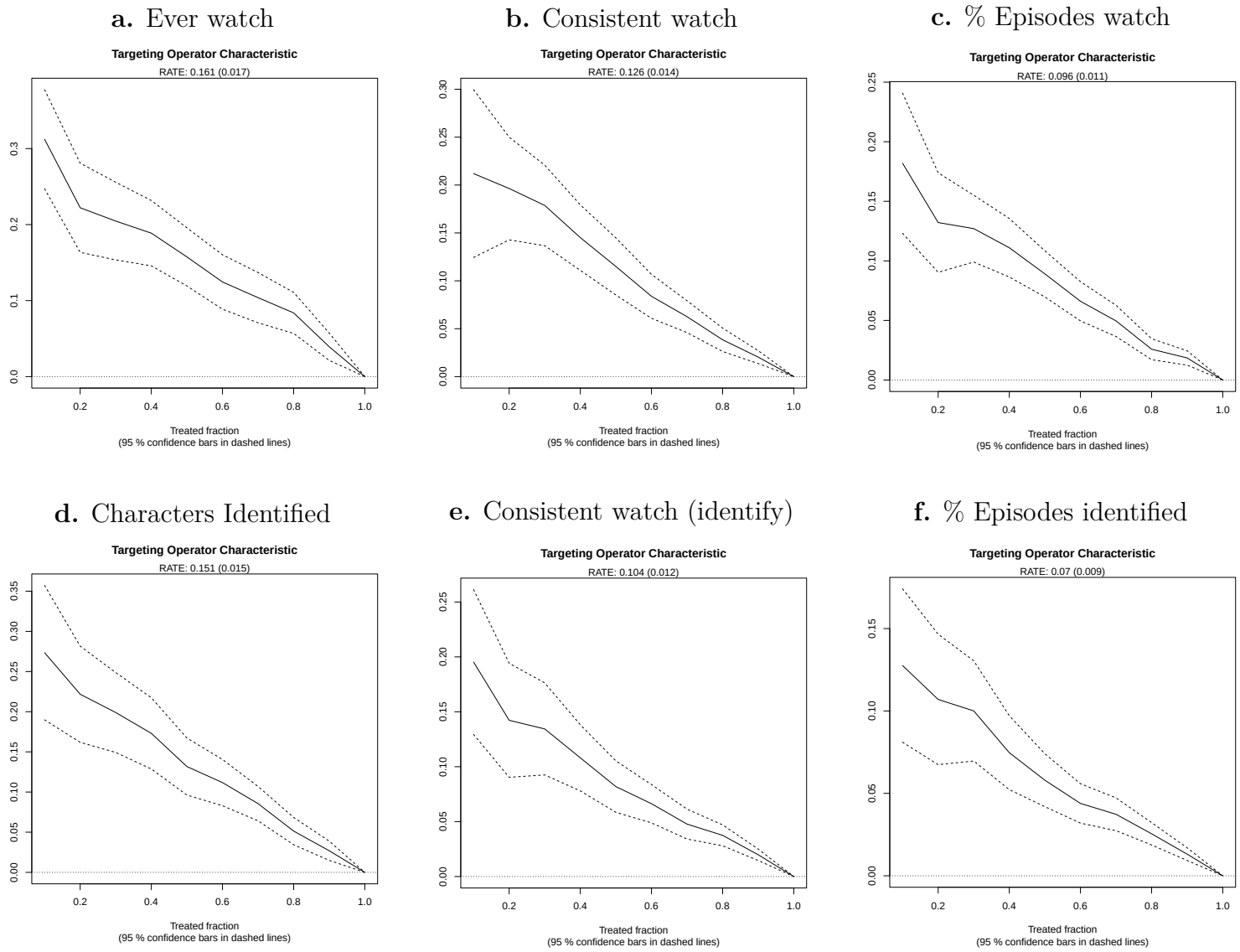


c. Gender Index



Notes: Figures present Targeting Operator Characteristics, showing how treatment effects vary across units ordered by predicted rank, based on a school-level clustered causal forest model (Wager and Athey, 2018). 95% confidence intervals indicated by dashed lines. Estimate for RATE and standard error in parenthesis at top of each figure calculated following Yadlowsky et al. (2024).

Figure A2: Assessing Heterogeneity in Compliance with Rank-Weighted Average Treatment Effects (RATE)



Notes: Figures present Targeting Operator Characteristics, showing how treatment effects vary across units ordered by predicted rank, based on a school-level clustered causal forest model (Wager and Athey, 2018). 95% confidence intervals indicated by dashed lines. Estimate for RATE and standard error in parenthesis at top of each figure calculated following Yadlowsky et al. (2024).

B Curriculum and Literacy Learning Objective by Episode of Nuzo & Namia

- **Episode 1:** Creating Images/Visualize - Readers create images using their senses to help draw conclusions, make predictions, interpret information, remember details, and improve comprehension. Picturing people, places, and events described in the story helps readers become emotionally involved.
- **Episode 2:** Skimming - This involves glancing quickly through a text to gain a general impression. Graphics, italics, and headings serve as useful cues when skimming.
- **Episode 3:** Questioning - Self-questioning encourages readers to think of questions before, during, and after reading to aid comprehension. Self-formulated questions provide a framework for active reading.
- **Episode 4:** Recalling - Remembering what was read by stating the main storyline, sequence of events, and characters in the story.
- **Episode 5:** Predicting - Prediction helps readers activate prior knowledge, combining it with new material in the text. Clues such as pictures, illustrations, subtitles, and plot hints guide predictions.
- **Episode 6:** Connecting - Connecting allows readers to make links between prior knowledge and new information in the text. Connections can be text-to-self, text-to-text, or text-to-world.
- **Episode 7:** Creating Images/Visualize - Repeated skill from Episode 1.
- **Episode 8:** Skimming - Repeated skill from Episode 2.
- **Episode 9:** Questioning - Repeated skill from Episode 3.
- **Episode 10:** Recalling - Repeated skill from Episode 4.
- **Episode 11:** Predicting - Repeated skill from Episode 5.
- **Episode 12:** Connecting - Repeated skill from Episode 6.
- **Episode 13:** Overview of learning objectives in Nuzo and Namia.

C Details on Gender Measures

The Gender Index is composed by three main items: stereotype knowledge, consistency and bias. Below, we provide details of how each of these measures are constructed.

- Gender stereotype knowledge: This item measures the level of stereotype knowledge that a child may have about a particular gender. Children's gender stereotype knowledge is another possible rigidity-related antecedent of gender attitudes. According to cognitive theories of gender development, the formation of gender identity is accompanied by the motivation to learn the associated features that distinguish one's gender from the other. This learning involves actively searching for and attending to gender-related information. The acquisition of this information forms the basis for descriptive gender norms— representations of girls and boys that are generalized to new individuals.

  To measure this, each child is asked whether they would typically see a boy, a girl, both or neither performing various actions (e.g., play with dolls, get in a fight) and exhibiting different traits (e.g., funny, caring). Following a mapping of which stereotypes are considered "male" or "female" in the literature (see Appendix Table A1), responses are coded with "1" if they follow the stereotype, "0" if they offer a neutral response, and "-1" if they follow a counter-stereotype. Positive numbers reflect a notion of counter-stereotype. These items are then aggregated into one score denoting the number of gender stereotypes that the child subscribes to, standardized and multiplied by minus one. Thus, a lower score means that the child has stronger (or more traditional) gender stereotype knowledge.

- Gender roles: This item measures what the child actually does or is inclined to do (activity/behavior) and the character traits they demonstrate. The sub-items within this are mapped exactly to the previous item on stereotypes and aims to elicit whether a child's behaviors and traits are consistent with their stereotype knowledge. The theoretical underpinning for this measure is to disentangle whether a child believes that their behaviors and traits need to be consistent with social messaging around gender roles. An important way Nuzo & Namia aims to impact gender attitudes is through demonstrating fluidity in gender roles that are antithetical to the stereotypes that may exist.

  Children are asked to rate on a scale of 1 to 4 how likely they are to perform certain activities or engage in behaviors (e.g., play with dolls) and asked the frequency to which they exhibit certain traits. Each item is coded as 1 if their answer follows the common

stereotype and 0 if not. The scores are averaged separately for consistency in behaviors and consistency in traits, each score is then standardized and multiplied by minus one. A negative score on this item indicates that the child has constancy in gender roles based on their stereotype knowledge.

- Gender Bias (In- and Out-group attitudes): This item aims to measure the perceptions and behaviors of a child towards their own gendered group and to those of their opposite gender. The theoretical underpinning of this item is- the higher the gender stereotype knowledge, the more consistent a child is with adhering to stereotypical gendered roles, and the more likely they are to demonstrate positive in-group attitudes and negative out-group attitudes.

  Children were asked two dimensions of attitudes: willingness to interact and actual interactions. For example: *Can you tell me how much you want to [or how often you] perform that activity on a scale of 1-3? 1 is Unwilling, 2 is somewhat willing, 3 is extremely willing: Join a group of [girls/boys] to study with after school.* Items are scored for in- and out- of group (i.e., same or opposite gender) where items are coded as 1 if they are somewhat willing or extremely willing. Then scores are averaged separately for in- and out- group, and for willingness and actual interactions (four measures), and multiplied by minus one. Finally, we calculate the difference between the in- versus out-group for willingness and for actual interactions. A negative value would indicate that the child is more biased towards their own gender in willingness or actual interactions.

D Adjusting for Non-Compliance

In this section we discuss estimates of the local average treatment effects of watching Nuzo & Namia. We modify equation 1 replacing the treatment indicator for a measure of compliance (e.g., ever watching the show, consistently watching the show, etc.), using the randomized encouragement design, that randomly *nudged* treated children to watch the show, as an instrumental variable for compliance. It is important to note that although the encouragement was randomized, there could be selection on which children were compliers. Therefore, estimates of the LATE would be reflecting the average effect for compliers in our sample.

Recall from the discussion in Section 2.5.1 that different measures of compliance capture different dimensions of compliance (intensive vs extensive) and have different weaknesses and strengths. For instance, self-reported "ever watch" the show could overstate take-up, thus provide a lower-bound estimate. On the other hand, the recall-based measure of consistently watching the show (identifying 6 episodes or more correctly) could be correlated with ability and understate the true take-up, thus providing an upper-bound estimate.

We present the 2SLS estimates in Appendix Table D1 for our primary outcomes in different Panels. Each column corresponds to a different compliance measure used as the instrumented endogenous variable. Because 2SLS assumes a linear first stage, using a control function estimator can improve efficiency by allowing for a nonlinear first stage (Wooldridge, 2015). Therefore, in Appendix Table D2 we replicate Table D1 but using a control function estimator for the binary endogenous variables instead. One drawback of using a control function is that it relies on a stronger distributional assumption for the first stage. Notably, coefficients are quite similar between Appendix Tables D1 and D2, and for some estimates we see large gains in precision when using the control function.

Estimates suggest large gains in comprehension from watching the show, though generally not robust to adjusting for multiple comparisons. For instance, ever watching the show increases comprehension by 19.5% of a standard deviation based on self-report compared to 25.2% of a standard deviation compared to the recall-based character identification question. Consistently watching the show is associated to stronger gains of 42.1% for self-reported "consistent watching" (6 episodes or more) or 104% for the recall-based correctly identifying 6 episodes or more. Although these estimates are all significant at the 10% level, they are not robust to accounting for multiple comparisons so should be interpreted with a grain of salt.

The estimates for curiosity and gender roles are similar in magnitude and robust to adjusting for multiple comparisons. Watching Nuzo & Namia increased curiosity by between 25.4-137% of a standard deviation depending on the measure used, and decreased beliefs

on gender roles by 23.9-128% of a standard deviation. These effects are precisely estimated (typically significant at the 5% level).

Table D1: 2SLS Estimates

	(1)	(2)	(3)	(4)	(5)	(6)
		Self-Report			Recall-Based	
Endogenous Variable:	Ever Watch	Consistent Watch	% Episodes Watch	Identify Characters	Consistent Watch (ID'd)	% Episodes Identified
Panel A: Dep.Vars. for Literacy						
Literacy Index	0.106	0.231	0.242	0.138	0.569	0.500
	(0.112)	(0.242)	(0.255)	(0.145)	(0.588)	(0.516)
Comprehension Index	0.195*	0.421*	0.443*	0.252*	1.040*	0.914*
	(0.117)	(0.250)	(0.262)	(0.149)	(0.611)	(0.530)
	[0.224]	*[0.224]*	*[0.224]*	*[0.224]*	*[0.224]*	*[0.224]*
Reading Index	0.013	0.028	0.029	0.017	0.068	0.060
	(0.099)	(0.216)	(0.226)	(0.129)	(0.531)	(0.466)
	[0.898]	*[0.898]*	*[0.898]*	*[0.898]*	*[0.898]*	*[0.898]*
Panel B: Dep.Vars. for Socio-Emotional Learning						
SEL Index	0.197**	0.428**	0.450**	0.256**	1.060*	0.928**
	(0.098)	(0.212)	(0.222)	(0.126)	(0.551)	(0.465)
Confidence score (z)	0.089	0.192	0.202	0.115	0.476	0.417
	(0.089)	(0.193)	(0.203)	(0.115)	(0.486)	(0.420)
	[0.433]	*[0.433]*	*[0.433]*	*[0.433]*	*[0.433]*	*[0.433]*
Curiosity score (z)	0.254**	0.552**	0.580**	0.330**	1.366**	1.198**
	(0.098)	(0.215)	(0.225)	(0.128)	(0.569)	(0.474)
	[0.050]	*[0.050]*	*[0.050]*	*[0.050]*	*[0.059]*	*[0.050]*
Panel C: Dep.Vars. for Gender Attitudes						
Gender Attitude Index	-0.119	-0.259	-0.272	-0.155	-0.640	-0.561
	(0.100)	(0.218)	(0.228)	(0.130)	(0.533)	(0.463)
Gender stereotype (z)	0.084	0.182	0.191	0.109	0.451	0.396
	(0.092)	(0.199)	(0.210)	(0.120)	(0.504)	(0.438)
	[0.433]	*[0.433]*	*[0.433]*	*[0.433]*	*[0.433]*	*[0.433]*
Gender roles (z)	-0.239**	-0.519**	-0.545**	-0.311**	-1.281**	-1.124***
	(0.093)	(0.204)	(0.212)	(0.122)	(0.514)	(0.429)
	[0.050]	*[0.050]*	*[0.050]*	*[0.050]*	*[0.050]*	*[0.050]*
Gender bias (z)	-0.243	-0.527	-0.554	-0.316	-1.304	-1.144
	(0.188)	(0.409)	(0.426)	(0.243)	(1.000)	(0.868)
	[0.346]	*[0.346]*	*[0.346]*	*[0.346]*	*[0.346]*	*[0.346]*
1st Stage F-Stats:	576.9	206.4	376.2	341.7	43.4	97.8

Notes: 2SLS estimates where the compliance measure (in columns) are the main endogenous variable, instrumented by the treatment indicator. All regressions include quadruple fixed effects, the baseline outcome as control and school controls. School-level controls (averages): share of girls, asset score, caregiver education, grade, reading score, HH size, TV watching time, and school size. Kleibergen-Paap first stage F-statistics presented. Q-values in brackets correspond to False Discovery Rate (FDR)-adjusted significance levels across all 7 primary outcome and 5 endogenous variables, following Benjamini et al. (2006). Cluster-robust standard errors at the school level in parenthesis. Significance: *** p<0.01, ** p<0.05, * p<0.1

Table D2: Control Function Estimates

	(1)	(2)	(3)	(4)
	Self-Report		Recall-Based	
Endogenous Variable:	Ever Watch	Consistent Watch	Identify Characters	Consistent Watch (ID'd)
Panel A: Dep.Vars. for Literacy				
Literacy Index	0.102	0.183	0.125	0.945***
	(0.112)	(0.212)	(0.134)	(0.356)
Comprehension Index	0.194*	0.403*	0.224	1.113***
	(0.118)	(0.216)	(0.138)	(0.396)
	[0.235]	*[0.176]*	*[0.235]*	*[0.092]*
Reading Index	0.006	-0.061	0.011	0.484
	(0.098)	(0.190)	(0.120)	(0.301)
	[0.956]	*[0.838]*	*[0.956]*	*[0.235]*
Panel B: Dep.Vars. for Socio-Emotional Learning				
SEL Index	0.181*	0.392**	0.238**	0.687*
	(0.096)	(0.196)	(0.118)	(0.360)
Confidence score (z)	0.077	0.227	0.136	0.411
	(0.088)	(0.177)	(0.106)	(0.300)
	[0.467]	*[0.330]*	*[0.330]*	*[0.326]*
Curiosity score (z)	0.238**	0.455**	0.278**	0.796**
	(0.097)	(0.198)	(0.121)	(0.379)
	[0.104]	*[0.104]*	*[0.104]*	*[0.114]*
Panel C: Dep.Vars. for Gender Attitudes				
Gender Attitude Index	-0.129	-0.178	-0.119	-0.422
	(0.101)	(0.197)	(0.126)	(0.363)
Gender stereotype (z)	0.084	0.153	0.087	-0.019
	(0.092)	(0.171)	(0.114)	(0.341)
	[0.467]	*[0.467]*	*[0.520]*	*[0.956]*
Gender roles (z)	-0.255***	-0.414**	-0.263**	-0.801**
	(0.093)	(0.191)	(0.119)	(0.348)
	[0.092]	*[0.109]*	*[0.109]*	*[0.104]*
Gender bias (z)	-0.256	-0.397	-0.240	-0.834
	(0.189)	(0.360)	(0.233)	(0.684)
	[0.326]	*[0.401]*	*[0.425]*	*[0.348]*
1st Stage F-Stats:	588.7	210.4	402.3	46.0

Notes: Control function (Wooldridge, 2015) estimates where the compliance measure (in columns) are the main endogenous variable. All regressions include quadruple fixed effects, the baseline outcome as control and school controls. School-level controls (averages): share of girls, asset score, caregiver education, grade, reading score, HH size, TV watching time, and school size. Kleibergen-Paap first stage F-statistics presented. Q-values in brackets correspond to False Discovery Rate (FDR)-adjusted significance levels across all 7 primary outcome and 5 endogenous variables, following Benjamini et al. (2006). Cluster-robust standard errors at the school level in parenthesis. Significance: *** p<0.01, ** p<0.05, * p<0.1

E Departures from the Pre-Analysis Plan

In the following section we provide details and justifications for any deviation from our pre-analysis plan below. We first note what section of the pre-analysis plan it corresponds to, followed by the corresponding details.

- Attrition: We present a more exhaustive analysis of attrition than what was detailed in the pre-analysis plan. The analysis on comparing characteristics of attritors and non-attritors in Appendix Table A1 was not part of the original pre-analysis plan. Because we do not find evidence of nonrandom attrition, we did not estimate Lee bounds.

- Compliance: For our measures of compliance we pre-registered self-reported measures for ever watch and consistent watch as reported. For recall-based we pre-registered character identification, and episode score. For consistency between self-report and recall-based measures, we added self-reported share of episodes watched and recall-based consistent watch (based on episode identification) using the same definition as the self-report variable. This addition does not change results, improves consistency and clarity, and if anything, makes estimates more conservative whenever adjusting for multiple hypothesis testing.

- Strata construction *(errata)*: the pre-analysis plan mentions 7 variables (averages at the school level) used to construct strata groups (quadruplets): share girls, asset score, grade, caregiver education, reading fluency score, size, tv watching. There was an eighth variable used to construct the strata which we missed including, that we now adding since it was used for constructing the strata: average household size.

- Outcome Definitions: The pre-analysis plan specifies defining a comprehension score rather than index. To maintain consistency across all outcomes, we calculate an index for comprehension following Anderson (2008) as done with all other outcomes. Results are not sensitive to defining the outcome as a simple average. Similarly, the pre-analysis plan defines the raw scores for confidence and curiosity. We standardize those for ease of interpretation.

- Outcome Definition: The pre-analysis plan calls for studying gender "in- and out-group attitudes". We redefined this as a single"gender bias", calculating it as the difference between in- and out- attitudes. This simplifies the interpretation of the variables in a single measure. This change does not materially change any results. Appendix Table E1 replicates Table 5 using the in- and out- attitudes instead of the single bias measure. Results are unchanged.

- Mechanisms: To study information as a mechanism, the pre-analysis plan mentioned: *"We can use items within the literacy tests to learn to what extent learning these strategies are driving results. In particular, there are questions that are mapped to different curriculum items covered by the show (e.g., recall, scanning, inference and prediction)."* In trying to code items for the comprehension test, it was hard to assign items to particular skills, as multiple skills may be relevant for any particular item. Therefore, we did not conduct this analysis.

- Discussion: The analysis presented in Appendix Table A19 was not part of our original pre-analysis plan (as explicitly mentioned in the text of Section 5). We added this analysis to try to better understand what could be driving the negative effect on gender roles.

Table E1: ITT Estimates - Primary Outcomes with in- vs out- attitudes

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
Panel A: Literacy	Literacy Index		Comprehension Index		Reading Index					
Treatment	0.043	0.049	0.080	0.090*	0.004	0.006				
	(0.055)	(0.052)	(0.057)	(0.054)	(0.050)	(0.046)				
q-value	-	-	0.428	0.257	0.942	0.898				
School Controls	No	Yes	No	Yes	No	Yes				
R^2	0.339	0.360	0.281	0.304	0.358	0.377				
Obs	3,669	3,669	3,669	3,669	3,669	3,669				
Panel B: SEL	SEL Index		Confidence		Curiosity					
Treatment	0.086*	0.091**	0.036	0.041	0.113**	0.117***				
	(0.046)	(0.044)	(0.042)	(0.041)	(0.046)	(0.044)				
q-value	-	-	0.631	0.575	0.055	0.045				
School Controls	No	Yes	No	Yes	No	Yes				
R^2	0.080	0.089	0.072	0.077	0.067	0.077				
Obs	3,669	3,669	3,669	3,669	3,669	3,669				
Panel C: Gender	Gender Index		Gender Stereotype		Gender Roles		Gender In-group		Gender Out-group	
Treatment	-0.052	-0.055	0.044	0.039	-0.108**	-0.110**	0.014	0.010	-0.034	-0.032
	(0.045)	(0.046)	(0.042)	(0.042)	(0.043)	(0.043)	(0.037)	(0.038)	(0.048)	(0.049)
q-value	-	-	0.587	0.575	0.055	0.045	0.797	0.898	0.648	0.677
School Controls	No	Yes	No	Yes	No	Yes	No	Yes	No	Yes
R^2	0.181	0.184	0.077	0.082	0.134	0.137	0.375	0.378	0.201	0.204
Obs	3,669	3,669	3,669	3,669	3,669	3,669	3,669	3,669	3,669	3,669

Notes: Each column presents estimates for an OLS regression corresponding to a different outcome or specification. All estimates include controls for the baseline dependent variable and quadruplet fixed effects. Column (2) adds School-level controls (averages): share of girls, asset score, caregiver education, grade, reading score, HH size, TV watching time, and school size. Column (3) presents q-values for adjusting for multiple comparisons across all 8 primary outcomes. Cluster-robust standard errors at the school level in parenthesis. Significance: *** p<0.01, ** p<0.05, * p<0.1

F Information Session Materials

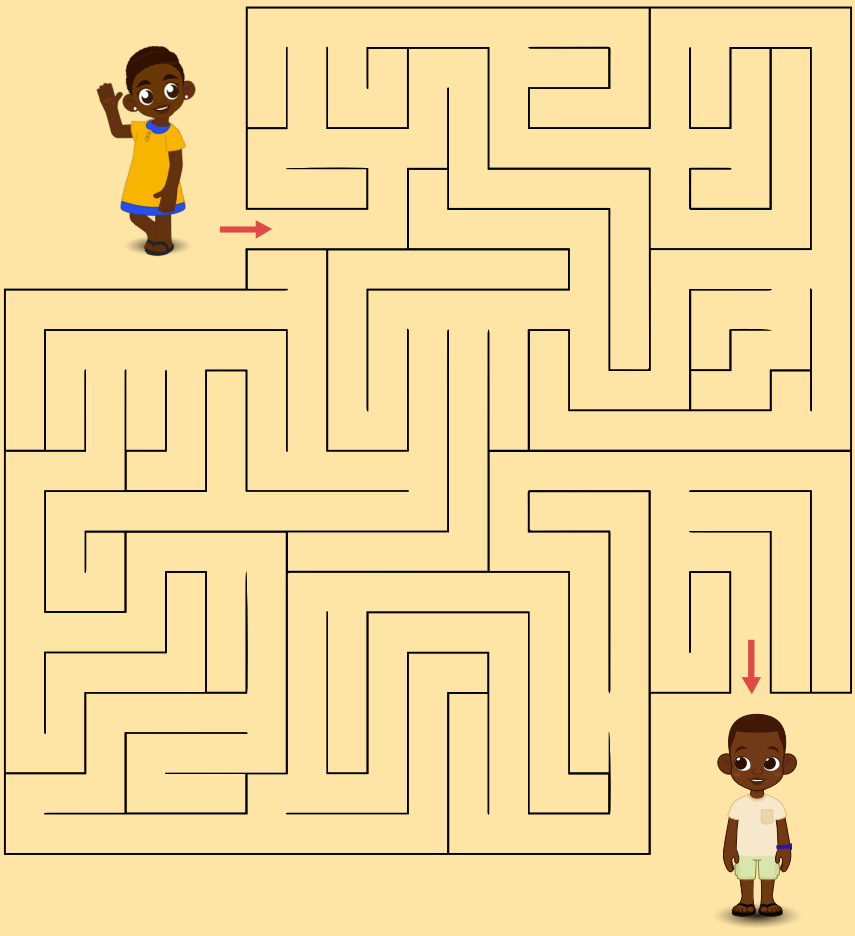
The following pages contain original and translated materials provided to parents at the information session. Note that the first poster was originally distributed in English and all other materials were originally in Kiswahili.

Nuzo and Namia

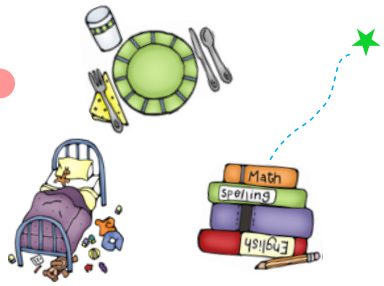
Time to have fun and learn

6.30pm – 7.00pm every Wednesday
and 10.30am – 11.00am every Saturday
on the Akili Kids Channel

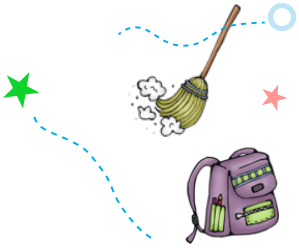
Help Namia find her way to Nuzo
in time for the next episode !



Akili Kids!



Nuzo and Namia



RATIBA YA KILA JUMAMOSI

Masaa	Kinachofanyika kabla ya kutazama “Nuzo and Namia”
6:00am ⇩ 10:30am	

Masaa	Kufurahia na kusoma na Nuzo na Namia kwenye Akili Kids	Kinachofanyika “Nuzo and Namia” inapotazamwa
10:30am ⇩ 11:00am		

Masaa	Kinachofanyika baada ya kutazama “Nuzo and Namia”
11:00am ⇩ 8:00pm	

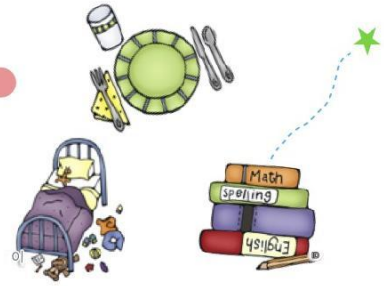
RATIBA YA KILA JUMATANO

Masaa	Kinachofanyika kabla ya kutazama “Nuzo and Namia”
4:00pm ⇩ 6:30pm	

Masaa	Kufurahia na kusoma na Nuzo na Namia kwenye Akili Kids	Kinachofanyika “Nuzo and Namia” inapotazamwa
6:30pm ⇩ 7:00pm		

Masaa	Kinachofanyika baada ya kutazama “Nuzo and Namia”
7:00pm ⇩ 9:00pm	





NUZO and Namia



EVERY SATURDAY SCHEDULE

10-0 10-0 10-0	What happens before watching "Nuzo and Namia"	
	Enjoy reading with Nuzo and Namia on Akili Kids	What happens when "Nuzo and Namia" is viewed
		
	What happens after watching "Nuzo and Namia"	

EVERY WEDNESDAY SCHEDULE

Hours		
4:00pm ⇩ 6:30pm	What happens before watching "Nuzo and Namia"	
Masaa 6:30pm ⇩ 7:00pm	Enjoy reading with Nuzo and Namia on Akili Kids	What happens when "Nuzo and Namia" is viewed
		
Masaa 7:00pm ⇩ 9:00pm	What happens after watching "Nuzo and Namia"	



Fomu ya Mzazi kujaza

93% ya wazazi/wanaotoa huduma kwa wazazi/walezi ambao wanawao huona kipindi cha Nuzo and Namia wamesema kwamba kipindi hiki ni bora!

Naweza kusema kipindi cha Nuzo and Namia kimeboresha maisha yangu kwa sababu watoto wangu wanapata elimu bila kulipia masomo ya ziada.

Baba ya watoto watatu walio na miaka 3, 5 na 7.

Mimi _____

na apa kwamba wanangu

_____

_____

_____

watakuwa wakitazama kipindi cha Nuzo and Namia kila Jumamosi saa 10.30 - 11.00 asubuhi katika Akili Kids Channel.

NUZO

Namia

by brain.

Parent Form to fill out

93% of parents/caregivers whose children watch the Nuzo and Namia show said that this show is the best!

> I can say that the Nuzo and Namia program has improved my life because my children are getting an education without paying for extra lessons.
>
> Father of three children aged 3, 5 and 7.



Me _____

and swear that my sons

_____

_____

_____

they will be watching the Nuzo and Namia program every Saturday at 10.30 - 11.00 am in Akili Kids Channel.

mind KIDS!

Ratiba ya Nuzo and Namia



Inafaa utazame Nuzo and Namia angalau **mara moja kila Jumamosi au Jumatano.**

Weka alama kwenye Jumamosi au Jumatano ambazo mtoto/watoto wako wametazama “Nuzo and Namia”.



Jumamosi					
Juni	Tarehe 24				
Julai	Tarehe 1	Tarehe 8	Tarehe 15	Tarehe 22	Tarehe 29
Agosti	Tarehe 5	Tarehe 12	Tarehe 19	Tarehe 26	
Septemba	Tarehe 2	Tarehe 9	Tarehe 16	Tarehe 23	Tarehe 30
Oktoba	Tarehe 7	Tarehe 14	Tarehe 21	Tarehe 28	
Novemba	Tarehe 4	Tarehe 11	Tarehe 18	Tarehe 25	

Jumatano					
Juni	Tarehe 21	Tarehe 28			
Julai	Tarehe 5	Tarehe 12	Tarehe 19	Tarehe 26	
Agosti	Tarehe 2	Tarehe 9	Tarehe 16	Tarehe 23	Tarehe 30
Septemba	Tarehe 6	Tarehe 13	Tarehe 20	Tarehe 27	
Oktoba	Tarehe 4	Tarehe 11	Tarehe 18	Tarehe 25	
Novemba	Tarehe 1	Tarehe 8	Tarehe 15	Tarehe 22	Tarehe 29

Schedule of

NUZO and Namia



You should watch Nuzo and Namia at least once every Saturday or Wednesday.

Mark the Saturdays or Wednesdays that your child/children have watched "Nuzo and Namia".



saturday					
June	The 24th				
July	Date 1	The 8th	The 15th	The 22nd	The 29th
August	The 5th	The 12th	The 19th	The 26th	
September	Date 2	The 9th	The 16th	The 23rd	The 30th
October	The 7th	The 14th	The 21st	The 28th	
November	The 4th	The 11th	The 18th	The 25th	

Wednesday					
June	The 21st	The 28th			
July	The 5th	The 12th	The 19th	The 26th	
August	Date 2	The 9th	The 16th	The 23rd	The 30th
September	The 6th	The 13th	The 20th	The 27th	
October	The 4th	The 11th	The 18th	The 25th	
November	Date 1	The 8th	The 15th	The 22nd	The 29th

Kipindi cha

Nuzo and Namia

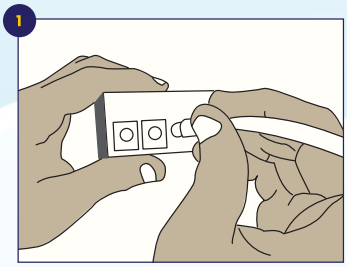
by ubongo

Kinaletwa ndani ya Akili Kids Channel kila Jumatano saa 6.30-7.00 jioni na Jumamosi saa 10.30-11.00 asubuhi.

Usipitwe na Kipindi cha Nuzo and Namia.

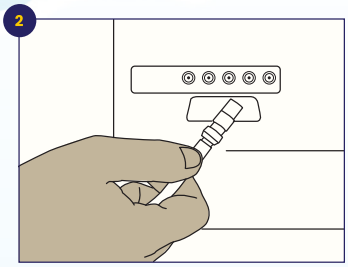
Fuata hatua zifuatazo ikiwa usajili wako umeisha, na hauna SMART TV ili wanako wasipitwe na Kipindi cha Nuzo and Namia.

1



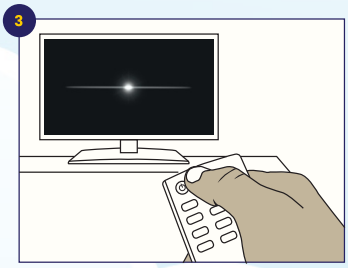
Ondoa sanduku la decoda. Sanduku hili linapatikana katikati ya antena na televisheni.

2



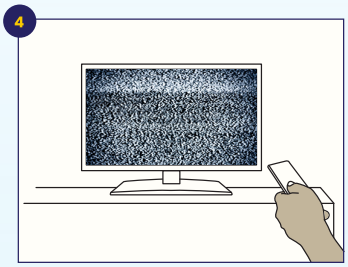
Wekelea antenna uliochukua kwenye sanduku kwa antenna ya televisheni

3



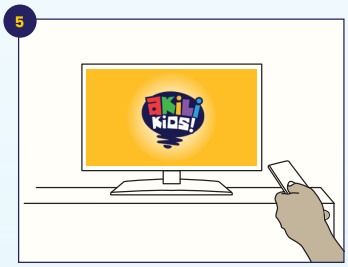
Zima televisheni na ungoje.

4



Fungua televisheni na tafuta chaneli za analogue.

5



Chagua Akili Kids.

Tulia na usome, utatue na ufurahie Nuzo and Namia na marafiki!

Akili Kids!

Period of

NUZO and Namia

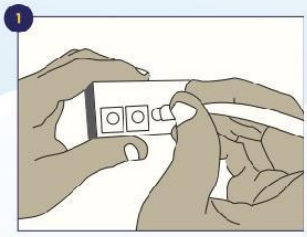
by brain.

It is introduced in Akili Kids Channel
every Wednesday at 6.30-7.00 pm
and Saturday at 10.30-11.00 am.

Don't miss the Nuzo and Namia Episode.

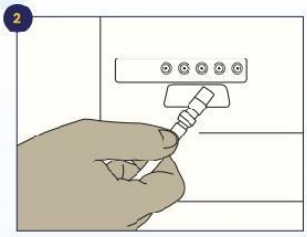
Follow the following steps if your subscription has expired, and you do not have a SMART TV so that your children do not miss the Nuzo and Namia Show.

1



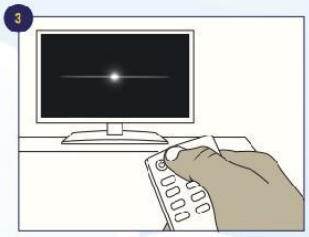
Remove the decoder box.
This box is located between the antenna and the television.

2



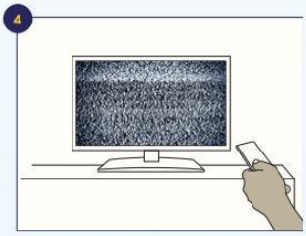
Attach the antenna that you took in the box to the television antenna

3



Turn off the television and wait.

4



Turn on the television and search for analogue channels.

5



Choose Smart Kids.

Sit back and read, solve and enjoy Nuzo and Namia with friends!

mind KIDS!