



The relationship between educational television and mathematics capability in Tanzania

Joe Watson , Sara Hennessy  and Anna Vignoles 

Joe Watson is a Research Assistant working with the EdTech Hub team in the Faculty of Education, University of Cambridge. He has recently completed a PhD focusing on educational television in sub-Saharan Africa. Sara Hennessy is a Reader in Teacher Development and Pedagogical Innovation in the University of Cambridge's Faculty of Education. She is also a Director of Research for the EdTech Hub which explores how technology can optimally be used to improve educational outcomes of marginalised learners. Anna Vignoles is a Professor of Education at the Faculty of Education, University of Cambridge, a Fellow of the British Academy, a trustee of the Nuffield Foundation and a member of the ESRC Council. Address for correspondence: Joe Watson, Faculty of Education, University of Cambridge, 184 Hills Road, Cambridge, CB2 8PQ, UK. Email: joe.mark.watson@gmail.com

Abstract

Previous studies have often demonstrated that educational television can have a positive effect on learning outcomes in low-income countries when delivered in controlled settings. However, existing research in low-resource contexts has scarcely considered the association between child outcomes and viewing in usual environments (ie, at their home, a friend's home or a relative's home). This lack of research is striking, as evidence from controlled settings might provide limited information on the effects of normal television exposure. This paper, therefore, investigates the relationship between normal exposure to a popular Tanzania-produced cartoon, *Ubongo Kids* and mathematics capability, as represented by plausible values derived from an item response theory model applied to children's test responses. Cross-sectional investigation of a sample of 38 682 Tanzanian children suggested normal educational television exposure to be significantly associated with mathematics capability, when controlling for age, sex, school enrolment, Kiswahili attainment and household fixed effects. While cross-sectional results are not necessarily causal, the findings in this paper broadly correspond with those from previous designs using repeated observations. What is more, considering association results alongside cost and viewership estimates suggests television-based interventions to be highly cost effective.

Keywords: cost effectiveness, educational television, mathematics, Tanzania

Introduction

Mathematics proficiency is limited in many low-income countries and mainland Tanzania (the focus of this paper) is no exception.¹ Only 49% of Tanzanian children aged 6 to 16 are able to correctly answer five mathematics items ranging in difficulty from number identification (easiest) to subtraction (hardest) (own calculations based on Uwezo, 2017 data).² Conversely, television technology is relatively prevalent. In Tanzania, 24% of children aged 6 to 16 live in households with televisions (ibid). Educational television shows could, therefore, provide a cost-efficient means

Practitioner Notes

What is already known about this topic

- National data for Tanzania suggests the country to have both low levels of mathematics proficiency and considerable access to television technology.
- Educational television-based interventions might, therefore, be worthy of consideration by educational policymakers, especially given that studies conducted in controlled settings indicate that television can deliver learning benefits.
- There is, however, little evidence concerning educational television exposure outside of controlled settings. Only one such study in a low-income context that concerned the association between learning outcomes and viewing among primary-age children has been identified.

What this paper adds

- This paper addresses this dearth of research by investigating the association between normal exposure to a Tanzanian cartoon, *Ubongo Kids* and mathematics capability, as derived from an item response theory model applied to the test responses of 38 682 children.
- Cross-sectional findings suggest the association between normal television exposure and mathematics capability to be significant.
- Further, a cost-effectiveness comparison with alternate interventions in comparable contexts indicates that educational television is highly cost-effective.

Implications for practice and/or policy

- The findings presented in this paper concerning educational television viewership in usual environments act to triangulate those from prior research conducted in controlled settings. As such, policymakers in low-income contexts now possess more convincing evidence on the potential influence of educational television interventions.
- Additionally, the cost-effectiveness comparison made suggests that educational television should be considered a viable option by policymakers seeking to address learning outcomes with limited resources.

of supporting the education of many Tanzanian children. As such, this study uses cross-sectional analysis of a national sample to investigate the association between normal (home-based) exposure to a popular Tanzanian mathematics-focused show and mathematics capability—represented by plausible values derived from child test responses (Section 3.1.2.1).

The mathematics-based show used as a vehicle for this investigation is *Ubongo Kids* (Figure 1). This programme is broadcast in Tanzania on Saturdays and Sundays (in both Kiswahili and English). 17% of Tanzanian children aged 6 to 16 report *Ubongo Kids* viewership (own calculations based on Uwezo, 2017 data). Additionally, the show is now broadcast in 40 African nations. Given the prevalence of *Ubongo Kids*, this paper should be of interest to those with and without a direct connection to the programme. Specifically, it is hoped that this paper provides information of relevance to researchers, practitioners and policymakers concerned with all forms of child-focused educational television in low-resource contexts.

Research questions

This paper addresses the following research questions:



Figure 1: Mathematics material on Ubongo Kids

- (Primary RQ) What is the association between normal exposure to educational television and mathematics capability in a low-income context?
- (Subsidiary RQ) How cost-effective is educational television in a low-income context?

Literature review

Educational television programmes targeted at children aged over 30 months have frequently been found to deliver learning benefits. The evidence on one such show, *Sesame Street*, “shows positive effects from studies of exposure to single episodes to studies of sustained, repeated viewing” (Anderson, Lavigne, & Hanson, 2013, p. 9). Indeed, this series has been identified to provide positive effects since its very first season (Ball & Bogatz, 1970). Numerous later studies have shown different programmes targeted at similar age groups to have positive effects on various school readiness indicators, such as literacy outcomes (for *Super Why*: Linebarger, McMennamin, & Wainwright, 2008) and problem-solving strategy (for *Blue’s Clues*: Crawley, Anderson, Wilder, Williams, & Santomero, 1999). Research also suggests that viewers benefited from educational shows intended for older child audiences, with science-based programmes including *Bill Nye the Science Guy* (appropriate for 7 years and above) and *3-2-1 Contact* (8 to 12 years) being found to further children’s understanding of fundamental scientific knowledge (Anderson *et al.*, 2013).

While studies have primarily occurred in high-income nations, child-focused studies in low-income countries have produced similar results. A meta-analysis conducted by Mares and Pan (2013) on 24 such studies investigating the effect of *Sesame Street*-based cartoons upon school readiness suggested an average effect size of 0.292. An paper included in the meta-analysis concerned the country of focus in this paper, Tanzania. There, a six-week intervention involving exposure to the Tanzanian version of *Sesame Street* was delivered via television, radio and print materials (Borzekowski & Macha, 2010). Randomised controlled trial (RCT) results suggested the intervention to promote various learning outcomes, health and hygiene behaviours and social and emotional development measures among pre-school children.

Borzekowski has also contributed to the small body of literature concerning educational television in low-resource contexts regarding programmes that were not (versions of) *Sesame Street*. The first of these contributions concerned an RCT involving 568 children (3 to 6 years) beginning

pre-primary school in Morogoro, Tanzania (Borzekowski, 2018). Children were divided into two groups, with a control group viewing 30 minutes of non-educational programmes for five days a week over four weeks and a treatment group watching *Akili and Me* at the same times. *Akili and Me*, like *Ubongo Kids*, is an animated cartoon produced in Tanzania by the educational media organisation, Ubongo, yet, *Akili and Me* is aimed at pre-school audiences. Regression models showed programme exposure to significantly improve children's scores in five of the seven foundational learning outcomes assessed in the study (Borzekowski, 2018). The second paper provided further evidence on the impact of *Akili and Me*, focusing on a slightly older sample in Rwanda (6 to 8 years; Borzekowski, Lando, Olsen, & Giffen, 2019). Here, the intervention led to significant gains in 8 of the 10 school-readiness competencies measured. The research design was again an RCT conducted over a short time period (two weeks) featuring one treatment and one control group.

Investigation into the implications of educational television exposure outside of controlled settings in low-income countries has been sparse. Since 2000, there have only been three published studies concerning the relationship between normal exposure and child educational outcomes in such nations (Lapinid *et al.*, 2017; Lee, 2009; Rimal, Figueroa, & Storey, 2013). Amongst these studies, only the research by Lapinid *et al.* concerns school-age children (Grade 3–6 in the Philippines).³ Indeed, the methodology employed by Lapinid *et al.* appears susceptible to criticism as insufficient allowance was made for differences between children in the treatment and control groups.

Methods

Information on the data for analysis

Uwezo data

The data for analysis were provided by Uwezo. Uwezo data give a nationally representative anonymised sample of Tanzanian children captured from household-based data collection. These data include survey responses from children and their caregivers as well as child test data on topics including basic Kiswahili and mathematics. Mathematics test items were made up of questions appropriate for children in Standard 2 (aged 8), covering concepts from number recognition to subtraction (see Figure 2). Child test questions were created by experts within the University of Dar es Salaam, the Tanzanian Institute of Education and local teachers (Uwezo, 2012). In 2017, children were asked one of three similar sets of mathematics questions, each of which includes one polytomous (main mathematics test) item and two binary (mathematics in everyday life) items. All mathematics questions featured competencies taught in *Ubongo Kids* episodes. However, *Ubongo Kids* content has frequently covered concepts not assessed by Uwezo (Section 4).

The child survey questions asked by Uwezo in 2017 included a child-reported measure of exposure to *Ubongo Kids*. This question was formulated by the lead author and submitted to Uwezo for inclusion. Per Uwezo requirements, the question was child-directed and included just two

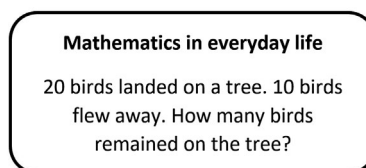


Figure 2: The subtraction-focused "mathematics in everyday life" question from "test set 2", translated from Kiswahili

response options. As such, children were asked whether they had watched *Ubongo Kids* in the last week, with responses recorded as “Yes” or “No/Do not know”. A short period (one week) was selected to combat potential difficulties in recalling exposure over a longer timeframe.

Data manipulation

A subset of the Uwezo, 2017 data set was created by selecting those children for whom information was available for any Uwezo mathematics item and all other variables used in the cross-section model (Section 3.2.1). These manipulations were carried out as the model employed was only informed by children with complete data. After this, children were retained from this subset only if there was information for another child in their household. This step was taken as the model used household-level fixed effects (requiring multiple children in each household), to control for within-household differences between children in the same households (Section 3.2.1). The final stage of data manipulation involved removing children from the data set if their assessor had not recorded their mathematics “test set”. This decision was made as the creation of plausible mathematics capability values for analysis through an IRT approach accounted for the varying difficulty level of items (see, Creation of plausible values), which were found to vary slightly between test sets.

These manipulations produced a final data set of 38 682 children with the following characteristics⁴:

- Age range and mean: 6 to 16 years, 10.64 years
- Proportion that are female: 49.92%
- Proportion that reported being enrolled in school: 88.03%
- Proportion that live in households with TVs: 15.75%
- Proportion that reported watching *Ubongo Kids* in the past week: 12.18%

Creation of plausible values

Mathematics values for use in analysis were created from child responses to Uwezo test items. The model presented in this paper uses capability estimates derived from the application of an item response theory (IRT) model applied to every mathematics item across all test sets.⁵ Using this IRT-based method means that capability estimates are influenced by all available mathematics information, while accounting for (amongst other things) the fact that certain items might be more difficult than others. It could be noted that prior analysis of Uwezo data has commonly featured transformation of child test results into a binary measure of success (see, for example, Alcott & Rose, 2016). However, this method was not selected as it would have removed nuance from a measure already limited in complexity.

To ensure that IRT was applicable, its assumptions were considered (using packages from the R statistical computing language). This involved applying confirmatory factor analysis (CFA) using the lavaan package to check unidimensionality (Rosseel, 2019), employing the mokken package’s check.monotonicity function to consider monotonicity (van der Ark, Straat & Koopman, 2018) and calculating Yen’s Q3 through the mirt package to investigate local independence (Chalmers, 2019). The application of CFA showed that the factor loading for each mathematics item on a single trait (termed, mathematics capability) was 0.860 or higher (with a typically adopted lower limit for this value being 0.3: Kline, 2011). Employing Yen’s Q3 showed that no item residuals possessed a positive correlation of greater than 0.2—the cut off at which Yen’s Q3 suggests items to be locally dependent (Chen & Thissen, 1997). There was also no apparent violation of manifest monotonicity by any item (established through application of the check.monotonicity function: van der Ark, 2007). This finding was supported descriptively by the shape of binary item curves (Figure 3), which suggest that the probability of responding positively to items increases with theta.

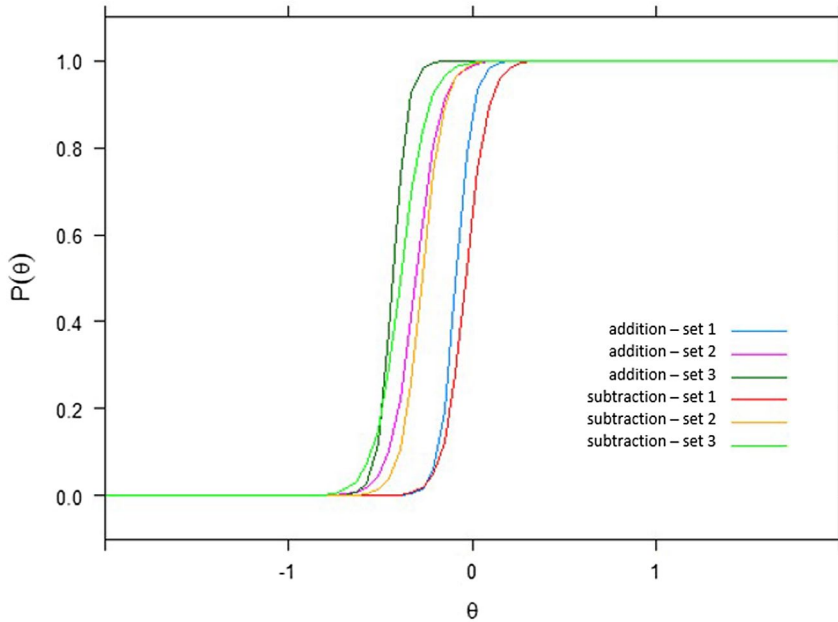


Figure 3: Item information plot for mathematics in everyday life items

Because the items comprising the Uwezo test were both polytomous and binary, either of the following IRT models could have been applied: a Graded Partial Credit Model (GPCM); or, a Rasch model (which is equivalent to a GPCM model but with item discrimination constrained to 1). Both models take account of varying item difficulty (and item category difficulty, for polytomous items), yet, only a GPCM accounts for varying item discrimination. As there was no theoretical justification for assuming item discrimination to be uniform, the GPCM model was selected. The relative applicability of a GPCM model over a Rasch model was supported by ANOVA comparison of log-likelihood information.

Investigation of the GPCM suggests that Uwezo mathematics items provide maximal information for test takers with low to average capability levels ($\theta = -1$ to 0). This is shown by plotting test and item information. Test and item information plots feature capability level (theta) on the x axis. The test information plot features total test information (at any given theta value) on the y axis. Conversely, item information plots feature probability on the y axis. Probability refers to the likelihood of answering any binary item correctly (for Figure 3) or the likelihood of a test taker achieving a particular test category (eg, “addition”: Figure 4).⁶ Item information plots support the assertion that maximal information is provided for test takers with low to average capability, as the midpoint ($P(\theta) = 0.5$) of item characteristic curves for binary items are located just below $\theta = 0$ (Figure 3) and the levels of the main mathematics test for children in “test set 1” are clustered around a similar value (Figure 4). Correspondingly, the shape of the test information curve peaks between $\theta = -1$ and $\theta = 0$ (Figure 5).

From the IRT model applied, multiple plausible values (PVs) were generated. PVs are random draws “from the distribution of scores that could be reasonably assigned to each individual” (Monseur & Adams, 2009, p. 6). As such, no individual PV should be considered to accurately

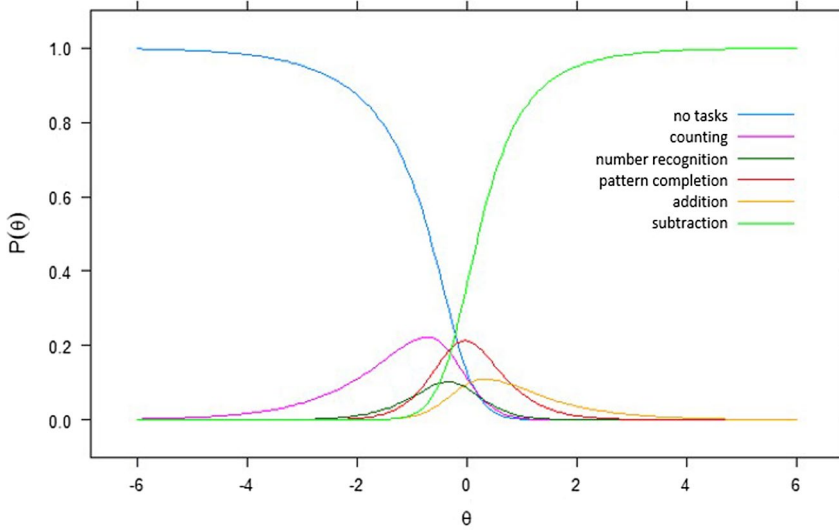


Figure 4: Item information plot for the main mathematics item from "test set 1"

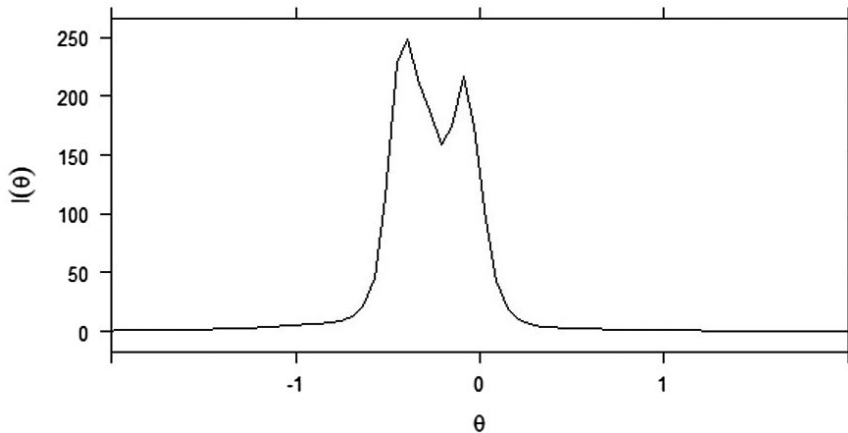


Figure 5: Test information plot

represent a child's capability. Instead, the PVs created give a range of mathematics capability values that each child might have. Using PVs represents a departure from previous IRT-based work featuring Uwezo data, where singular IRT-derived point estimates were produced for each child (Jones, 2017). This paper employs PVs as opposed to point estimates, as the latter would ultimately have led to biased standard errors on regression coefficients (OECD, 2009). 20 PVs were created per child. The selection of this number followed previous simulation-based studies (Luo & Dimitrov, 2019). Trialling also suggested 20 PVs to be sufficient, as a comparison of regression coefficient results against those produced using singular point estimates (from expected a posteriori scores) did not show any substantive differences.⁷

Approach

Those studies cited in the literature review featured numerous approaches that could potentially have been applied to investigate educational television in Tanzania. However, not all methods were applicable. The longstanding availability and popularity of *Ubongo Kids* in Tanzania makes it challenging to identify children with no prior exposure, thereby hindering a random allocation of treatment (which featured in other studies including Borzekowski & Macha, 2010). Indeed, a design in which treatment “under hypothetical and contrived conditions” is randomly allocated might not provide the most appropriate means of investigating normal exposure to television (Nagin & Sampson, 2019, p. 140). To support this point, it could be noted that Borzekowski’s (2018, p. 55) Tanzania-based study involved showing children five episodes over four weeks “with children watching each episode for four days in succession”. It is improbable that viewing in usual environments would follow this pattern. As such, another approach was used to carry out investigation into the association between normal educational television exposure and mathematics capability: cross-sectional investigation of Uwezo data for Tanzania.

Cross-sectional regression model

In the cross-sectional model used to examine the relationship between mathematics capability and *Ubongo Kids* exposure, child capability values were regressed on multiple fixed independent variables. These variables were self-reported viewership, Kiswahili attainment, child age and child school enrolment status. The proxy for exposure was provided by the self-reported viewership measure described in Section 3.1.1. The remaining variables on which mathematics PVs were regressed provided controls.

A Kiswahili attainment measure was employed to control for children’s non-mathematics outcomes. This comprised a binary measure of “success”, dependent on whether children answered all components of the sole Uwezo Kiswahili test item correctly (Alcott & Rose, 2016). Including this measure as a control would have acted to limit bias resulting from children with better outcomes (across all subjects) or higher levels of motivation being more likely to watch *Ubongo Kids*. All remaining controls were selected in accordance with previous analyses of Uwezo data, which suggested learning outcomes to be positively related to a child’s (current) self-reported school enrolment status (enrolled or not enrolled: Alcott & Rose, 2015), age (Jones, 2017) and, to a small degree, (female) gender (Jones & Schipper, 2015).

Further, all within-household (unobserved and observed) differences between children were controlled for using household-level fixed effects. Such differences could have included (but were not limited to) whether a child’s mother had attended school and household wealth, both of which have been found to be related to child outcomes in Uwezo data analysis (see: Alcott & Rose, 2016; Mugo, Ruto, Nakabugo, & Mgalla, 2015, respectively). Lastly, the model accounted for the nested structure of Uwezo data (described by Uwezo, 2016) by applying clustered standard errors at the enumeration area level and district level.

Analysis was repeated for all 20 of each child’s plausible capability values, using the `withPV` function from the R package, `mitools` (Lumley, 2019). The `Mlcombine` function from `mitools` was then used to combine the results of these analyses, giving a singular set of coefficients and standard errors for each independent variable. Lastly, p-values (which are not provided through analysis using `mitools`) were estimated by taking the median p-value for each coefficient from all regression analyses applied to each set of PVs (Eekhout, Van De Wiel, & Heymans, 2017).

The cross-section model is presented in the equation below:

$$\begin{aligned}
 (\text{MathsCap}_{ih} - \overline{\text{MathsCap}_h}) = & \beta_0 + \beta_1 \left(\text{UbRep}_{ih} - \overline{\text{UbRep}_h} \right) + \beta_2 \left(k_{ih} - \bar{k}_h \right) + \beta_3 \left(\text{Age}_{ih} - \overline{\text{Age}_h} \right) \\
 & + \beta_4 \left(\text{Sex}_{ih} - \overline{\text{Sex}_h} \right) + \beta_5 \left(\text{Enr}_{ih} - \overline{\text{Enr}_h} \right) + (\varepsilon_{ih} - \bar{\varepsilon}_h)
 \end{aligned}$$

In this formula, *MathsCap* represents mathematics capability, *UbRep* denotes self-reported exposure to *Ubongo Kids*, *k* represents Kiswahili attainment, *Age* refers to child age, *Sex* refers to child's sex and *Enr* concerns a child's current school enrolment status. Subscript *ih* denotes information for an individual, *i*, in a specific household, *h*. The bar accent (eg, \bar{k}) is used to show a sample average. In all cases, this is employed at the household level (eg, \bar{k}_h), thereby referring to the mean result of a household in the sample.

Before estimating this equation, normality was investigated among all non-binary variables: mathematics capability and age.⁸ This examination was conducted as the validity of parametric tests, including regression, requires variables to be normally distributed (Field, Miles, & Field, 2012). Normality was gauged by assessing kurtosis and skewness as well as comparing mean and median values. Mean and median results were similar for both variables, which supports the assumption that distributions were normal. Further, results from kurtosis and skewness calculations fell within broadly accepted boundaries (-2 to 2) (calculated using the *e1701* package in R).

Results

Findings from the analysis are presented in Table 1. These findings suggest normal (home-based) *Ubongo Kids* exposure—like Kiswahili attainment, age and school enrolment—to be positively and significantly related to mathematics capability.⁹ Further, the size of this relationship appears notable. The coefficient for the association between mathematics capability and *Ubongo Kids* exposure (0.130) is greater than that for an additional year in age (0.059), although these coefficients are not significantly different from one another.

It is acknowledged that findings for the treatment variable should be treated with caution. This is partly because the regression model did not control for child school-type or pre-school attendance, due to high levels of missingness among these variables in the Uwezo data. This omission could be important, as both school-type and pre-school attendance have been found to be related to learning outcomes in prior Uwezo analysis (Alcott & Rose, 2016; Bietenbeck, Ericsson, & Wamalwa, 2017). Additionally, the inability to follow children between different time points due to their anonymisation in Uwezo data meant that longitudinal analysis was not possible. Because of these omissions, the presence of a causal relationship between *Ubongo Kids* exposure and mathematics capability cannot be inferred.

Table 1: Coefficients for the associations between independent variables and mathematics capability

Independent variable	Coefficient	Std. error	p-value
<i>Ubongo Kids</i> viewership	0.130	0.049	.002**
Kiswahili attainment	0.933	0.028	.000***
age	0.059	0.003	.000***
school enrolment	0.197	0.030	.000***
sex (female)	0.002	0.012	.705

Note. The above table presents the coefficient estimate for each independent variable, along with its standard error and p-value. On the right-hand side of the table, symbols are provided to indicate the level at which each coefficient estimate is significant. These are: "***" for 0.001, "**" for 0.01, "*" for 0.05, "." for 0.1 and "" for 1.

The model does, however, have numerous strengths. It was applied to a large data set derived from the Uwezo, 2017 survey, which was structured to be representative of Tanzania as a whole. A large amount of unobserved heterogeneity was accounted for, through the application of household-level fixed effects. Further, the model controlled for key child characteristics for which there was within-household variance. In doing so, the model included a proxy for non-mathematics outcomes. The employment of this control would have acted to limit the bias that would have arisen if children with higher levels of motivation or multi-subject capability were more likely to have watched *Ubongo Kids*.

What is more, the consideration of measurement error concerning mathematics capability and *Ubongo Kids* exposure supports the idea that results may have underestimated the association between both concepts. This measure was highly unlikely to have captured all educational outcomes that *Ubongo Kids* might have affected. That is, the programme could have been associated with outcomes in mathematics topics beyond the Uwezo assessment or in other subject areas such as English and science. Each of these assertions can be supported by detailing key features of *Ubongo Kids*' format and content:

- *Ubongo Kids* airs in both English and Kiswahili.
- Some episodes have focused on science-based topics, such as "Battle of the body parts" which primarily concerns anatomy.
- Mathematics-focused episodes like "Fish billionaire" have covered topics including division and basic fractions, both of which are more complex mathematics topics than any assessed in Uwezo's, 2017 assessment (where subtraction was the hardest concept).

Should the programme have been related to outcomes in English, science or mathematics topics outside the scope of Uwezo assessment, the treatment coefficient would have only partially reflected *Ubongo Kids*' association with all affected educational outcomes.

It was also probable that the measurement of exposure was susceptible to bias, which would have decreased the likelihood of finding exposure to be significantly related to mathematics capability. Child-reported viewership has been found to provide useful information on exposure (Rimal *et al.*, 2013), yet, the limited complexity of the exposure proxy employed in this paper might have introduced imprecision. Responses to the binary viewership question in the 2017 Uwezo survey, "did you watch *Ubongo Kids* in the last week?" (translated from Kiswahili), did not permit differentiation between more and less frequent viewers. Variation in the frequency of exposure amongst viewers would have increased standard errors, thereby decreasing the probability of finding a significant association.

Additionally, viewership question responses might have failed to capture any prior *Ubongo Kids* exposure in various instances. This would have occurred if, for example, the Uwezo survey respondent: did not recognise the name of the show (as they referred to it by another name); failed to recall recent exposure; or, had simply been exposed at any time before (but not during) the week preceding assessment. In any of these events, children who *had been exposed* to *Ubongo Kids* at some point would simply have been *treated as not exposed*. Conversely, it is feasible that children misreported (positive) viewership of *Ubongo Kids*, which might have happened if children were confused about which show they were being questioned about; or, falsely reported viewership in the belief that this was a "correct" or "desired" response. In such instances, children who *had not benefited* from viewing *Ubongo Kids* would have been *treated as having been exposed*. Any situation in which children who had or had not benefited from viewing *Ubongo Kids* were mis-categorised would have contributed to underestimation of the treatment coefficient.

Lastly, it is recognised that the treatment coefficient identified in the paper is smaller than some found in comparable studies. This includes those studies that controlled for pre-intervention outcomes. To demonstrate this, the identified coefficient for *Ubongo Kids* exposure is depicted alongside available mathematics-focused results from a selection of studies referenced above (Figure 6). In this graphic, the coefficients from Tanzania-based studies concerning *Sesame Street* (Borzekowski & Macha, 2010) and *Akili and Me* (Borzekowski, 2018) as well as the Rwanda-focused research concerning *Akili and Me* (Borzekowski *et al.*, 2019) were produced through RCTs. Coefficients concerning a version of *Sesame Street* in Bangladesh show the effect of normal exposure while controlling for lagged outcomes (Lee, 2009). Additionally, the study in Egypt gives results from cross-sectional research concerning another *Sesame Street* variant (Rimal *et al.*, 2013). All included results were significant at $p = .05$ or below.¹⁰ Considering the coefficient for *Ubongo Kids* against those obtained in prior studies provides little evidence to suggest that the findings in this paper are unrepresentative of the effect of educational television. Indeed, the coefficient for this study is far less than for all those found in selected RCTs.

This section has provided evidence to suggest that exposure to *Ubongo Kids* is positively related to mathematics capability. This was shown through the application of a regression model that controlled for key child characteristics including non-mathematics outcomes and household fixed effects. Indeed, exploration of the measures of mathematics capability and television exposure used in the model suggested that the identified treatment coefficient was likely biased downwards. While the inability to employ a measure of pre-intervention capability still denies any claim of causality, results correspond with those from longitudinal investigation into the effects of normal exposure to a pre-primary show (Lee, 2009). Further, RCTs conducted in Tanzania and Rwanda concerning different educational television interventions have found far greater treatment coefficients (Borzekowski, 2018; Borzekowski *et al.*, 2019; Borzekowski & Macha, 2010). Policymakers should, therefore, be aware that educational television interventions could provide a viable means of targeting mathematics outcomes in low-income contexts. The following section explores the implications of results by considering *Ubongo Kids*' cost effectiveness.

Cost-effectiveness analysis

Cost-effectiveness components

Cost-effectiveness analysis (CEA) permits determination of “the least cost approach to meeting such educational objectives as ... raising [attainment]” or the relative gains in attainment achieved by differing interventions for a given cost (Levin, 1988, p. 52), with this paper focusing on the latter. Comparing the relative gains from investments suffers from both imprecision and limited nuance. Cost-effectiveness comparisons do not account for differences in the location of programmes' beneficiaries. Additionally, a number of the values used to create CEA results are inherently challenging to quantify (as demonstrated by subsequent discussion of user costs). However, the employment of CEA still provides a valuable indication of educational television's cost effectiveness.

The cost-effectiveness analysis (CEA) presented below follows the approach advocated by the Abdul Latif Jameel Poverty Action Lab (J-PAL: Abdul Latif, 2014), as outlined in Dhaliwal, Duflo, Glennerster, and Tulloch (2012). J-PAL material forms a substantial portion of educational research concerning cost effectiveness, which is limited albeit growing (Levin & McEwan, 2001). As CEA comparisons require the use of a common approach, using the J-PAL method, therefore, permits consideration against a range of interventions. Numerous projects examined using the J-PAL method have suggested positive cost effectiveness (with standard deviation gains per \$100 spent ranging from 0.06 to 118.34, based on impact findings published in: Baird, McIntosh, &

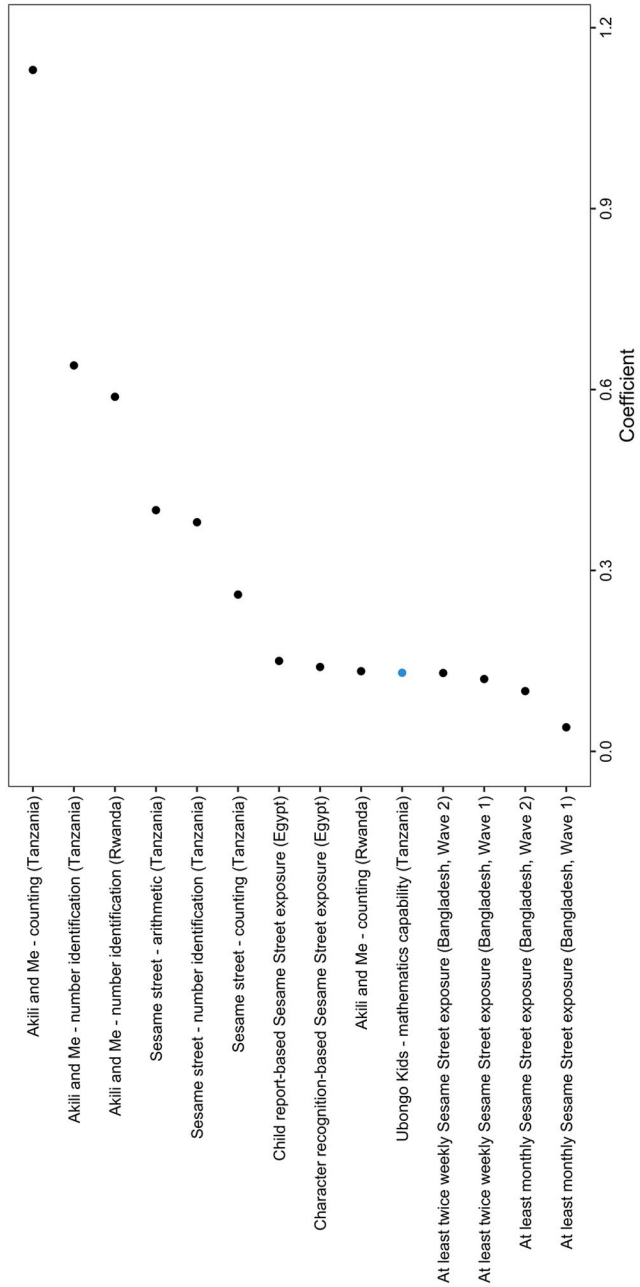


Figure 6: Comparison of treatment coefficients

Özler, 2011; Nguyen, 2008, respectively). However, to our knowledge there have been no cost-effectiveness evaluations of educational television interventions from 2000 onwards (using the J-PAL approach or otherwise).

Calculating CEA through the J-PAL method requires information on a programme's cost, influence and number of beneficiaries, to populate the following equation:

$$\text{Change in standard deviations per \$100 spent} = 100 / (\text{Cost/Influence} \times \text{Beneficiaries})$$

In this equation, "Cost" refers to the total programme cost, "Influence" refers to the standard deviation gain among individuals that is attributable to the programme and "Beneficiaries" refers to the total number of children who have benefited from programme exposure.

The following subsections give information on the estimation of *Ubongo Kids*' number of beneficiaries, costs and influence. These estimates are intended to relate to *Ubongo Kids*' ongoing operations, as opposed to its activities since inception (following the approach taken to consider another continuing intervention: Sabates, Alcott, Rose, & Delprato, 2018). To consider ongoing operations, a duration period of one year (2017) was selected. Using this duration period allowed the end point of "ongoing activities" to fall immediately after Uwezo data collection (December 2017). Additionally, it was intended that the start point of this period (January 2017) provided a compromise between upwards and downwards biases on the final CEA result, which are subsequently explored in discussion of the intervention's "Number of beneficiaries" and "Influence".

Number of beneficiaries

An estimate for the number of *Ubongo Kids* beneficiaries was produced using United Nations (UN) population estimates (accessed 2019) and Uwezo, 2017 data. Uwezo data provided the only means of gauging the proportion of children (6 to 16 years) exposed to *Ubongo Kids* from nationally representative child-level data. This was established by calculating the weighted mean of responses in the Uwezo, 2017 data set to a question regarding recent viewership, which suggested that 17% of children had watched the show in the past week.

To estimate the number of child viewers for 2017, the percentage of viewers was multiplied by the total number of children aged 6 to 16 in Tanzania (as of 2017) from UN estimates for Tanzania. UN estimates were created using censuses and other official reports, with adjustments for underenumeration (<https://population.un.org/wpp/DataSources/834>). These sources gave an approximation of the total Tanzanian population aged 6 to 16 years for 2017: 15,842,916 children. Multiplying this number by the percentage of children of the same age who reported *Ubongo Kids* viewership gave a figure of 2 643 250 viewers.

Adopting a programme duration period of 2017 makes it probable that this calculation is an underestimate. This assertion is based on similar reasoning to that employed when noting that the exposure proxy used in the cross-section model would have failed to reflect viewership among children who had watched *Ubongo Kids* before but not during the week leading up to Uwezo assessment (Section 4). Drawing on the child-reported viewership measure when estimating the number of beneficiaries would have meant that numerous children sampled by Uwezo who had benefited from viewing *Ubongo Kids* at some point in 2017 could (accurately) have reported that no viewership occurred in the week prior to Uwezo assessment. For example, data from a sampled child who had viewed *Ubongo Kids* consistently during the first half of 2017 (only) would not have increased the estimated viewership figure.

Costs

Cost estimates were derived from figures submitted by Ubongo in accordance with the J-PAL basic costing template (available at: <https://www.povertyactionlab.org/research-resources/cost-effectiveness>). This template facilitated the establishment of total programme costs, by providing a framework for obtaining cost data in the following categories:

- Programme administration and staff costs
- Targeting costs
- Staff training
- Implementation and programme material costs
- Monitoring costs
- Participant training
- User costs
- Averted costs

The figures submitted for multiple cost categories comprising the J-PAL template had a value of zero. These categories were “Participant training”, “User costs”, “Averted costs” and “Targeting costs”. “Averted costs” were zero because the intervention was unlikely to have prevented any costs being incurred by beneficiaries or other education providers. Additionally, “Participant training” costs were nil because no specific training is required for viewing *Ubongo Kids*. What is more, “Targeting costs” were considered to be zero as Ubongo incurred no costs in 2017 associated with raising awareness about the intervention among potential Tanzanian beneficiaries.

The case for “User costs” being equal to zero was more nuanced than for any other category. Viewers devoted time resources to watching *Ubongo Kids* and could only have accessed the show using a media platform that likely cost a non-zero amount. However, an estimate of zero was still considered most appropriate. It would be farfetched to assume that the cost of such devices was incurred solely to watch *Ubongo Kids*. Additionally, findings from interviews with programme viewers (reported elsewhere: Watson, 2019) suggested that the opportunity cost of watching *Ubongo Kids* was often negligible. Viewers stated that time spent watching *Ubongo Kids* was frequently at the expense of playing games or viewing different television programmes.

Ubongo provided non-zero estimates for all remaining cost categories. These estimates reflected costs such as office rent (a component of “Administration and staff costs”), the fees spent on external voice actors (an “Implementation and material” cost), the cost of staff courses (a “Staff training” cost) and the subscription fee paid to an SMS-based viewership survey (a “Monitoring” cost). As *Ubongo Kids* broadcast in Kenya, Rwanda, Uganda and Tanzania in 2017, costs that could not be disaggregated by country were multiplied by the percentage of programme viewers located in Tanzania (23%).¹¹ All costs were summed and translated to 2011 US dollars, to facilitate CEA comparison with J-PAL estimates (which used the same currency and year).

Influence

The final component of the CEA formula considered is *Ubongo Kids*’ influence on learning outcomes. Influence is partially derived from the relationship coefficient between *Ubongo Kids* and mathematics capability (0.130; Table 1). This coefficient is multiplied by the standard deviation of the mathematics capability variable (1.001).¹² This calculation gives a point estimate for influence per individual of 0.131 standard deviation gains. Upper and lower bound estimates at the 90% confidence interval were also produced.

Introducing a figure derived from the exposure coefficient to the estimation of *Ubongo Kids*’ ongoing cost-effectiveness creates biases. These biases are likely to be multi-directional. Cost-effectiveness results could be overestimated, should those who reported recent viewership during

Uwezo assessment (in December 2017) have benefited from exposure prior to January 2017 (when treatment was assumed to have commenced for the purposes of this CEA calculation). Conversely, children who reported recent viewership during Uwezo assessment might not have benefited from the intervention throughout 2017 (as, for example, they first watched *Ubongo Kids* in December 2017). Downwards bias could also be exerted by underestimation of the exposure coefficient (see Section 4).

Cost-effectiveness findings

Cost effectiveness was calculated as follows: intervention costs were divided by the estimated influence per beneficiary multiplied by the total number of beneficiaries. This calculation was performed using a point estimate for the influence of *Ubongo Kids* and the upper and lower bound estimates of this figure (see Table 2).

CEA estimates for *Ubongo Kids* are plotted against those for all other interventions assessed under the J-PAL approach (for which positive results based on significant treatment findings were publicly available: Figure 7). Following the numbering in Figure 7, estimates [2] to [14] and [17] are taken from J-PAL: Abdul Latif (2014), who compiled information from the following studies: Baird *et al.*, (2011) [17]; Kremer, Miguel, and Thornton (2009) [13]; Burde and Linden (2013) [10]; Nguyen (2008) [2]; Glewwe, Kremer, and Moulin (2009) [7]; Abeberese, Kumler, and Linden (2014) [14]; Banerjee, Cole, Duflo, and Linden (2007) [8] and [12]; Kremer, Duflo, and Dupas (2011) [3] and [11]; Glewwe, Ilias, and Kremer (2010) [6]; Duflo, Hanna, and Ryan (2012) [9]; and, Pradhan *et al.*, (2011) [4] and [5]. Additionally, the Camfed estimate [15] is obtained from Sabates *et al.* (2018) and The Partnership Schools finding [16] from Romero *et al.* (2017).

The CEA estimate for *Ubongo Kids* suggests that the intervention can deliver greater child learning outcome benefits per \$100 spent than any other programme assessed using the J-PAL method. This finding was not primarily attributable to *Ubongo Kids*' influence on outcomes (0.131 standard deviations), which was lower than all but one of the studies against which its CEA result was ultimately compared (ranging from 0.130 to 0.588). Similarly, *Ubongo Kids*' CEA results were not greatly benefited by the programme's relatively high total cost (\$25 481.12), which was only exceeded by five of the J-PAL-assessed studies in Figure 7 (which ranged from \$409.67 to \$232 406 38).¹³

Instead, *Ubongo Kids*' strong CEA performance was predominantly attributable to its scale, which meant that the programme achieved a high total impact and low per-child cost (under \$0.01 per child). The estimated number of programme beneficiaries (2 643 250) far exceeds that of any other intervention achieving a positive CEA result (759 to 143 199 beneficiaries). Indeed, the scale of *Ubongo Kids*' intervention permits cost-effectiveness results unrealistic for programmes not delivered through mass media. This point acts to highlight a limitation of the CEA comparison made.

The CEA comparison in this paper does not account for the fact that certain forms of intervention cannot feasibly produce cost-effectiveness results approaching those of *Ubongo Kids*. Such

Table 2: CEA estimates

Estimate type	Cost per additional sd	Additional sd per \$100
point estimate	0.07	1,354.28
upper bound	0.05	2,194.31
lower bound	0.19	514.26

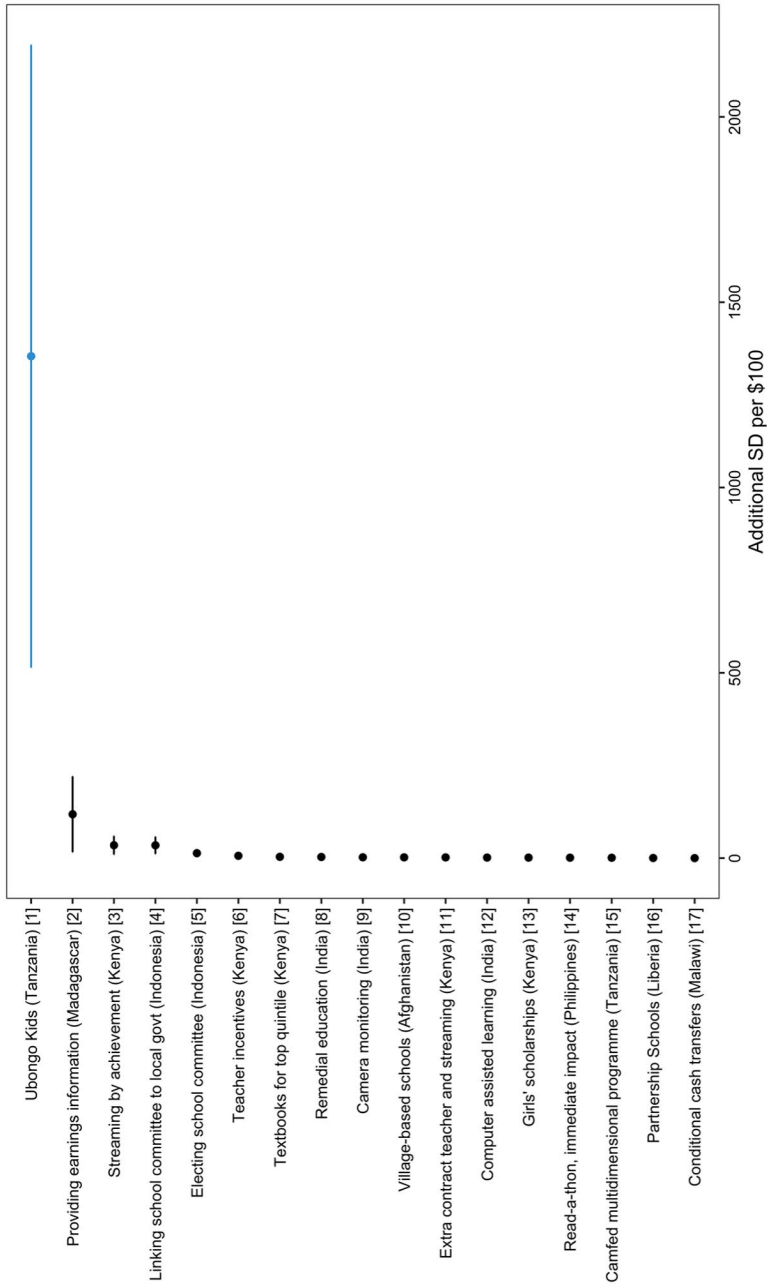


Figure 7: CEA comparison

interventions might include school-based interventions, where per-child costs are inevitably higher. The projects compared might also have differed in terms of context, population age and education level (Sabates *et al.*, 2018). What is more, the measure of *Ubongo Kids'* influence used in CEA calculations stemmed from cross-sectional analysis. This method of analysis might be considered less precise than the longitudinal approaches used to assess those interventions against which *Ubongo Kids* was compared. Nevertheless, *Ubongo Kids'* CEA result remains notable: this finding substantially exceeds that of other interventions, with it even possible that this result was underestimated (Section 5.1).

Conclusion

This paper has explored the relationship between mathematics capability among Tanzanian children and normal (home-based) exposure to the educational cartoon, *Ubongo Kids*. Findings suggest that exposure is positively associated with mathematics capability. It is acknowledged that the approach used in this paper did not control for mathematics capability at earlier time points and, therefore, cannot necessarily identify a causal impact. However, the model applied did control for household fixed effects in addition to numerous child characteristics, including non-mathematics outcomes. Indeed, exploration into the measures of mathematics capability and exposure employed in this model suggested that the association between *Ubongo Kids* exposure and mathematics capability could even have been underestimated. The claim that the identified association was not biased upwards was also supported by comparing the coefficient for *Ubongo Kids* exposure against those from other studies in similar contexts (Section 4).

The findings from this paper provide a valuable addition to previous research on educational television in low-income countries. This previous research has scarcely considered viewing outside of controlled settings (especially with regards to primary-age children: Section 2) and featured no cost-effectiveness evaluations in recent years (Section 5.1). Addressing these research gaps has important implications. The finding that mathematics capability and television exposure are significantly related pertains to normal *Ubongo Kids* exposure—something received by millions of Tanzanian children. Further, estimating *Ubongo Kids'* influence, cost and number of beneficiaries permitted a CEA comparison that suggested the educational television intervention under consideration to have been highly cost effective. This indicates that educational television initiatives are worthy of consideration by policy makers operating in resource-constrained contexts.

Acknowledgements

We are extremely grateful to those at the Tanzania office for Uwezo, for including a question submitted by the lead author in their 2017 survey and providing advance access to 2017 data pending public its release. We are also appreciative of all the respondents who participated in the Uwezo 2017 data collection. Additionally, we would like to thank those at Ubongo for allowing *Ubongo Kids* to be the focus of this study and providing the cost estimates used in this paper. Lastly, we acknowledge the support of the DFID-funded EdTech Hub (<https://edtechhub.org/>) in the writing phase of this work and the helpful input from anonymous peer reviewers to an earlier version of the paper.

Statements on open data, ethics and conflict of interest

The data in this study was provided by Uwezo Tanzania following a request for data access made by the lead author. Readers or future researchers seeking to access the data used in this paper

could also request data access through communication with Uwezo Tanzania or wait for Uwezo's upcoming publication of the data set.

Formal ethical approval for the study described in this paper was obtained through multiple agreements. A Tanzanian research visa was granted to the lead author following approval of the design presented in this paper by the Tanzanian Commission for Science and Technology (COSTECH). Additionally, compliance with the ethical guidelines set forth by the University of Cambridge's Faculty of Education was inferred by the Faculty's acceptance of a completed ethical clearance form, submitted before research was conducted. It should also be recognised that all decisions were made in accordance with the BERA Ethical Guidelines for Educational Research (BERA, 2011).

The authors confirm that there is no conflict of interest in this study.

Notes

¹The country of Tanzania encompasses both mainland Tanzania and Zanzibar. However, all findings and descriptive statistics presented in this paper concern mainland Tanzania only.

²There is significant variation in outcomes amongst children of each age (in years) for which test results are available in the Uwezo 2017 data (6 to 16 years). For this reason, it is considered feasible to use only the test administered by Uwezo to consider achievement across a broad age range. Access to the Uwezo 2017 data – before its public release – was provided by Uwezo following a request from the lead author.

³In the Philippines, children in Grade 3 are typically 8 to 9 years old and children in Grade 6 are typically 11 to 12 years. The article by Lapinid and colleagues (2017) does not, however, report the ages.

⁴These figures were calculated for those in the final dataset without accounting for sample weighting. As such, the statistics for recent *Ubongo Kids* viewership and household TV ownership presented in Section 3.1.2 differ to those presented elsewhere (Section 1 and Section 5.1).

⁵Estimates derived from IRT models are typically considered to concern an unobservable "ability" trait (denoted as θ). However, the mathematics estimates in this paper are instead presented as measures of "capability". Employment of the term, capability, follows identification that the term it acknowledges, ability, is problematic in educational contexts (Hart *et al.*, 2004).

⁶Figure 4 presents trace lines only for the main mathematics item from test set 1 to promote readability. Trace lines for test sets 2 and 3 appear very similar.

⁷PVs were not conditioned on background variables. Employing likely determinants of capability to support PV creation could enhance PV accuracy, yet might lead to endogeneity if these determinants are themselves employed as independent variables in subsequent regression analysis (Jerrim *et al.*, 2017).

⁸An assumption of unidimensional IRT is that the underlying trait distribution is normal (despite IRT being robust to some level of non-normality: Cotton & Baker, 2019). Despite this, PV estimates were still investigated to ensure they possessed characteristics suggesting normality. All checks for normality concerning mathematics capability were carried out using the first set of PVs.

⁹These results did not differ substantively from those identified through an alternate design – employed as an initial test of robustness – that featured a differently formulated dependent variable (following Delprato & Sabates, 2015). In this alternate model, the dependent variable was a binary measure of "success" in the main Uwezo mathematics test (as employed by Alcott & Rose, 2016).

¹⁰Figure 6 does not include one of the three exposure measures employed in Rimal *et al.*'s (2013) study (caregivers' reports of children's viewership), as results are insignificant.

¹¹This figure was established using SMS-based survey data from GeoPoll, an organisation that offers research services in low-income contexts (www.geopoll.com).

¹²The standard deviation of mathematics PVs was calculated by taking the mean standard deviation of each PV.

¹³This range excludes values for the Camfed multidimensional programme (Tanzania) and Partnership Schools (Liberia) programme, as total cost information was not available in either case.

References

- Abeberese, A. B., Kumler, T. J., & Linden, L. L. (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.
- Alcott, B., & Rose, P. (2015). *How can education systems become equitable by 2030?* (DFID Think Pieces). DFID. http://uis.unesco.org/sites/default/files/documents/how-can-education-systems-become-equitable-by-2030-learning-and-equity_pauline-rose_benjamin-alcott_heart_2015-en.pdf.
- Alcott, B., & Rose, P. (2016). Does private schooling narrow wealth inequalities in learning outcomes? Evidence from East Africa. *Oxford Review of Education*, 42(5), 495–510.
- Anderson, D. R., Lavigne, H. J., & Hanson, K. G. (2013). The educational impact of television. In A. N. Valdivia (Ed.), *The international encyclopedia of media studies* (pp. 1–22). Hoboken: Wiley-Blackwell.
- van der Ark, L. A. (2007). Mokken scale analysis in R. *Journal of Statistical Software*, 20(1), 1–19.
- Baird, S., McIntosh, C., & Özler, B. (2011). Cash or condition? Evidence from a cash transfer experiment. *The Quarterly Journal of Economics*, 126(4), 1709–1753.
- Ball, S., & Bogatz, G. A. (1970). *The first year of Sesame Street: An evaluation. Final report, volume III of Volumes* (PR-70-15; ETS Program Report). New York: Children's Television Workshop.
- Banerjee, A. V., Cole, S., Duflo, E., & Linden, L. (2007). Remedying education: Evidence from two randomized experiments in India. *The Quarterly Journal of Economics*, 122(3), 1235–1264.
- Bietenbeck, J., Ericsson, S., & Wamalwa, F. M. (2017). *Preschool attendance, school progression, and cognitive skills in East Africa* (Working Paper IZA DP No. 11212; Discussion Paper Series, pp. 1–38). Bonn: IZA Institute of Labor Economics.
- Borzekowski, D. (2018). A quasi-experiment examining the impact of educational cartoons on Tanzanian children. *Journal of Applied Developmental Psychology*, 54, 53–59.
- Borzekowski, D., Lando, A. L., Olsen, S. H., & Giffen, L. (2019). The impact of an educational media intervention to support children's early learning in Rwanda. *International Journal of Early Childhood*, 51(1), 1–18.
- Borzekowski, D., & Macha, J. E. (2010). The role of Kilimani Sesame in the healthy development of Tanzanian preschool children. *Journal of Applied Developmental Psychology*, 31(4), 298–305.
- Burde, D., & Linden, L. (2013). The effects of village-based schools: Evidence from a randomized controlled trial in Afghanistan. *American Economic Journal: Applied Economics*, 5, 27–40.
- Chen, W. H., & Thissen, D. (1997). Local dependence indexes for item pairs using item response theory. *Journal of Educational and Behavioral Statistics*, 22(3), 265–289.
- Cotton, J., & Baker, S. T. (2019). A data mining and item response mixture modeling method to retrospectively measure Diagnostic and Statistical Manual of Mental Disorders-5 attention deficit hyperactivity disorder in the 1970 British Cohort Study. *International Journal of Methods in Psychiatric Research*, 28(1), 1–13.
- Crawley, A. M., Anderson, D. R., Wilder, A., Williams, M., & Santomero, A. (1999). Effects of repeated exposures to a single episode of the television program Blue's Clues on the viewing behaviors and comprehension of preschool children. *Journal of Educational Psychology*, 91(4), 630–637.
- Delprato, M., & Sabates, R. (2015). A multilevel analysis of late entry in Nigeria. *International Journal of Research & Method in Education*, 38(1), 83–99.
- Dhaliwal, I., Duflo, E., Glennerster, R., & Tulloch, C. (2012). *Comparative cost-effectiveness analysis to inform policy in developing countries*. https://www.povertyactionlab.org/sites/default/files/publications/CEA%20in%20Education%202013.01.29_0.pdf.
- Duflo, E., Hanna, R., & Ryan, S. P. (2012). Incentives work: Getting teachers to come to school. *American Economic Review*, 102(4), 1241–1278.

- Eekhout, I., Van De Wiel, M. A., & Heymans, M. W. (2017). Methods for significance testing of categorical covariates in logistic regression models after multiple imputation: Power and applicability analysis. *BMC Medical Research Methodology*, 17(1), 129.
- Field, A., Miles, J., & Field, Z. (2012). *Discovering statistics using R*. London: SAGE Publications.
- Glewwe, P., Ilias, N., & Kremer, M. (2010). Teacher incentives. *American Economic Journal: Applied Economics*, 2(3), 205–227.
- Glewwe, P., Kremer, M., & Moulin, S. (2009). Many children left behind? Textbooks and test scores in Kenya. *American Economic Journal: Applied Economics*, 1(1), 112–135.
- Hart, S., Dixon, A., Drummond, M., & McIntyre, D. (2004). *Learning without limits*. Maidenhead: Open University Press.
- Jerrim, J., Lopez-Agudo, L. A., Marcenaro-Gutierrez, O. D., & Shure, N. (2017). What happens when econometrics and psychometrics collide? An example using the PISA data. *Economics of Education Review*, 61, 51–58.
- Jones, S. (2017). *Do schools in low income countries really produce so little learning?* University of Copenhagen. https://economics.handels.gu.se/digitalAssets/1643/1643680_17.-sam-jones-2017-02-20.schoolprod.v2.pdf.
- Jones, S., & Schipper, Y. (2015). Does family background matter for learning in East Africa? *Africa Education Review*, 12(1), 7–27.
- J-PAL: Abdul Latif Jameel Poverty Action Lab. (2014). *Cost-effectiveness comparisons: Full workbook*. J-PAL. <https://www.povertyactionlab.org/research-resources/cost-effectiveness>.
- Kline, R. B. (2011). *Principles and practice of structural equation modeling*, 3rd ed. New York: Guilford Publications.
- Kremer, M., Duflo, E., & Dupas, P. (2011). Peer effects, teacher incentives, and the impact of tracking. *American Economic Review*, 101(5), 1739–1774.
- Kremer, M., Miguel, E., & Thornton, R. (2009). Incentives to learn. *The Review of Economics and Statistics*, 91(3), 437–456.
- Lapinid, M., Gustilo, L., Magno, C., Barrot, J., Gabinete, M., & Anito, J. (2017). For the Nth Time, does educational technology really make a difference? - A large-scale investigation of the effects of educational TV on academic achievement. *Advanced Science Letters*, 23(2), 759–763.
- Lee, J. H. (2009). *Sisimpur's reach and educational impact: Evidence from a national longitudinal survey of a Sesame Street project in Bangladesh*. Sesame Workshop. <http://downloads.cdn.sesame.org/sw/SWorg/documents/ACPR+Monograph+Bangladesh+FINAL.pdf>
- Levin, H. M. (1988). Cost-effectiveness and educational policy. *Educational Evaluation and Policy Analysis*, 10(1), 51–69.
- Levin, H. M., & McEwan, P. J. (2001). *Cost-effectiveness analysis: Methods and applications* (Vol. 4). London: SAGE Publications.
- Linebarger, D. L., McMenamin, K., & Wainwright, D. K. (2008). *Summative evaluation of Super WHY!*. University of Pennsylvania. http://www-tc.pbskids.org/read/files/SuperWHY_Research_View.pdf.
- Luo, Y., & Dimitrov, D. M. (2019). A short note on obtaining point estimates of the IRT ability parameter with MCMC estimation in Mplus: How many plausible values are needed? *Educational and Psychological Measurement*, 79(2), 272–287.
- Mares, M.-L., & Pan, Z. (2013). Effects of Sesame Street: A meta-analysis of children's learning in 15 countries. *Journal of Applied Developmental Psychology*, 34(3), 140–151.
- Monseur, C., & Adams, R. (2009). Plausible values: How to deal with their limitations. *Journal of Applied Measurement*, 10(3), 320–334.
- Mugo, J. K., Ruto, S. J., Nakabugo, M. G., & Mgalla, Z. (2015). A call to learning focus in East Africa: Uwezo's measurement of learning in Kenya, Tanzania and Uganda. *Africa Education Review*, 12(1), 48–66.
- Nagin, D. S., & Sampson, R. J. (2019). The real gold standard: Measuring counterfactual worlds that matter most to social science and policy. *Annual Review of Criminology*, 2, 123–145.
- Nguyen, T. (2008). Information, role models and perceived returns to education: Experimental evidence from Madagascar. *Unpublished Manuscript*, 6. <http://xxpt.ynjgy.com/resource/data/20091115/U/>

- MIT20091115040/OcwWeb/Economics/14-771Fall-2008/Readings/PaperJM%20TRANG%20NGUYEN%2022jan08.pdf.
- OECD. (2009). *PISA data analysis manual: SPSS 2nd. ed.* OECD Publishing. https://www.oecd-ilibrary.org/education/pisa-data-analysis-manual-spss-second-edition_9789264056275-en.
- Pradhan, M., Suryadarma, D., Beatty, A., Wong, M., Alishjabana, A., Gaduh, A., & Artha, R. P. (2011). *Improving educational quality through enhancing community participation: Results from a randomized field experiment in Indonesia* (Working Paper No. 5795). The World Bank. <https://openknowledge.worldbank.org/bitstream/handle/10986/3559/WPS5795.pdf?sequence=1>.
- Rimal, R. N., Figueroa, M. E., & Storey, J. D. (2013). Character recognition as an alternate measure of television exposure among children: Findings from the Alam Simsim Program in Egypt. *Journal of Health Communication*, 18(5), 594–609.
- Romero, M., Sandefur, J., & Sandholtz, W. A. (2017). *Can outsourcing improve Liberia's schools? Preliminary results from year one of a three-year randomized evaluation of partnership schools for Liberia* (Working Paper No. 462). Centre for Global Development. <https://www.cgdev.org/publication/partnership-schools-for-liberia>.
- Sabates, R., Alcott, B., Rose, P., & Delprato, M. (2018). *Cost-effectiveness with equity: Raising learning for marginalised girls through Camfed's programme in Tanzania* (Policy Brief Policy Paper No. 18/2). REAL Centre, University of Cambridge.
- United Nations. (n.d.). *United Nations (UN) population estimates*. Retrieved 20 October 2019, from <https://population.un.org/wpp/Download/Standard/Interpolated/>.
- Uwezo. (2012). *Are our children learning: Report for 2011*. Twaweza. http://www.uwezo.net/wp-content/uploads/2012/08/TZ_2011_AnnualAssessment-Report.pdf.
- Uwezo. (2016). *Data cleaning protocol*. Twaweza. <https://www.twaweza.org/uploads/files/TZ15-Data%20cleaning%20and%20usage%20NOTES%202015.pdf>.
- Uwezo. (2017). *Data for Tanzania (pending public release) [Data file]*. Tanzania: Author.
- Watson, J. (2019). *The relationship between educational television and mathematics capability in Tanzania* [Doctoral Dissertation], University of Cambridge. <https://www.repository.cam.ac.uk/handle/1810/307624>.

R packages

- van der Ark, A., Straat, J. H., & Koopman, L. (2018). *mokken*. Conducts Mokken Scale Analysis. <https://cran.r-project.org/web/packages/mokken/mokken.pdf>.
- Chalmers, P. (2019). *mirt*. Multidimensional Item Response Theory. <https://cran.r-project.org/web/packages/mirt/mirt.pdf>.
- Lumley, T. (2019). *mitools*: Tools for Multiple Imputation of Missing Data. <https://cran.r-project.org/web/packages/mitools/mitools.pdf>.
- Rosseel, Y. (2019). *lavaan*: Latent Variable Analysis. <https://cran.r-project.org/web/packages/lavaan/lavaan.pdf>.