University of St Andrews

School of Economics and Finance Discussion Papers

Fostering Early Childhood Development in Low-Resource Communities: Evidence from a Group-Based Parenting Intervention in Tanzania

Margaret Leighton, Anitha Martine and Julius Massaga

School of Economics and Finance Discussion Paper No. 2204
21 Apr 2022
JEL Classification: J13, I21, I25, I28, J18
Keywords: early child development, child care policy, parenting, impact evaluation, Tanzania

School of Economics and Finance
Online Discussion Paper Series
issn 2055-303X
http://ideas.repec.org/s/san/wpecon.html
info: econ@st-andrews.ac.uk

Fostering Early Childhood Development in Low-Resource Communities: Evidence from a Group-Based Parenting Intervention in Tanzania*

Margaret Leighton†, Anitha Martine‡, Julius Massaga‡

January 31, 2022

Abstract

Group-based parent training programmes present an affordable means to influence the early experiences of children at scale. This paper reports evidence on the effectiveness of one such early child development programme piloted in rural Tanzania. The core of the intervention is an 8-10 week caregiver training course led by local facilitators, build around early stimulation and nurturing care. After two years of implementation, the intervention led to improvements in the development of 3-year olds of 0.26 standard deviations. Detailed data on caregivers indicates that these improvements are due to changes in the type and

*The authors gratefully acknowledge the team from Save the Children International Tanzania, Save the Children UK and ADP Mbozi who designed and carried out the intervention on which this research is based, in particular John Tobongo, Emily Weiss, Kirsten Mucyo, Richard Germond, and Celine Sieu. The project on which the paper is based would not have been possible without the participation of the Government of Tanzania and the caregivers and children in Songwe region. The authors thank Shyamal Chowdhury, Deborah Cobb-Clark, Thomas Dohmen, David Escamilla-Guerrero, Jakob Henning and David Jaeger, as well as seminar participants at the Institute for Social and Economic Research (University of Essex), the University of Glasgow and the University of St Andrews for helpful comments and discussions. Research assistance by Lawrence Ho and Himangshu Kumar is gratefully acknowledged. Data collection was approved by the Tanzania Commission for Science and Technology, with ethical clearance granted by the National Institute of Medical Research (NIMR/HQ/R.8a/Vol.IX/2670; NIMR/HQ/R.8a/Vol.IX/3228). This research has been approved by the St Andrews' University Teaching and Research Ethics Committee (EC15104). The Tuwekeze Pamoja intervention is funded by Comic Relief. Margaret Leighton gratefully acknowledges research funding from the Scottish Funding Council Global Challenges Research Fund and the Royal Society of Edinburgh Research Re-Boot (Covid-19 Impact) Research Grants. Conflict of interest statement: Margaret Leighton has worked as a paid consultant to Save the Children on other projects. Contribution statement: ML data analysis, interpretation and writing of the manuscript. AM and ML: contributed to design of the study and data collection plan. JM: principal investigator for data collection.

†University of St Andrews. Corresponding author: mal22@st-andrews.ac.uk; Castlecliffe, The Scores, St Andrews, KY16 9AR, UK

‡Save the Children International Tanzania

frequency of caregiver-child interactions, as well as the quality of play materials in the home.

Keywords: early child development, child care policy, parenting, impact evaluation, Tanzania
JEL codes: J13 I21 I25 I28 J18

1 Introduction

The science of early child development has established that the early years, from conception to age five, are a critical time where adverse conditions, or positive interventions, can have a life-long impact (Doyle et al. (2009); Black et al. (2017)). Recent estimates suggest that as many as 43% of children in this critical age range living in low-income and middle-income countries are at risk of failing to meet their developmental potential, with rates as high as 66% in Sub-Saharan Africa (Lu et al. (2016)). Such set-backs early in the life cycle carry high long-term costs, including earnings losses later in life; however, a growing body of evidence has identified effective and scaleable intervention approaches that can prevent the establishment of developmental delays (Richter et al. (2017)).

A growing body of research has responded to the urgent need for such programmes to be adapted, tested and rolled-out at scale in areas where children are at high risk. Rigorous studies have investigated the efficacy of interventions which feature enhancing the quality of centre-based pre-school care (e.g. Ozler et al. (2018), Chujan and Kilenthong (2021), Andrew et al. (2019)), home visiting (e.g. Attanasio et al. (2014); Andrew et al. (2018), Grantham-McGregor et al. (2020)) and parent education (e.g. Ozler et al. (2018), Grantham-McGregor et al. (2020)). These studies have demonstrated the potential of such interventions, but have also highlighted many challenges.

In 2017, Save the Children initiated a pilot study of a caregiver-focussed early child development intervention in rural Tanzania. Called Tuwekeze Pamoja, the intervention was designed with scale-up in mind, and sought to promote early child development by supporting caregivers' development of early stimulation and nurturing care practices within the home. Group-based caregiver training sessions were delivered by local facilitators through 8-10 weekly meetings. The pilot study included a second treatment arm, which received the main intervention plus additional programming designed to tackle violence in the home.

This included five additional group sessions, as well as community-focussed events.

This paper evaluates the effect of the Tuwekeze Pamoja intervention on early child development over the first three years of life. Following a panel of child-caregiver pairs, first surveyed when the child was 4-12 months and followed up two years later, the evaluation seeks to answer three related questions. First, what impact did the intervention have on child development over these two years? Second, what further impact can be attributed to the additional package of violence-reduction programming? And finally, through what channels of home environment and caregiver practices did the intervention operate, if any?

Before the intervention, a set of control areas was chosen to match the treatment areas as closely as possible: a sample of child-caregiver pairs was then surveyed in each area. Treated and control observations were similar at baseline, both in terms of the primary outcome variable and on objective demographic criteria. A comparison of these two groups shows that the intervention had a substantial positive effect on early child development. Intention to treat estimates show an improvement in child development of 0.26 standard deviations, while instrumental variables estimates of the effect of the treatment on the treated are considerably higher at 0.51 standard deviations. Improvements of a similar magnitude are found across four domains of child development: motor, cognitive, language and socio-emotional. These estimates imply a marginal cost effectiveness of between £36-£83 per standard deviation increase in child development.

While the core intervention shows substantial improvements in child development, there are no differences in the treatment effects across the two arms of the study. This suggests that the additional package of violence reduction programming either led to minimal changes in the use of violence in the home, or that any such changes did not translate into improvements in child development on the time scale of this study. These additional investments have so far failed to demonstrate any value-added.

Detailed data on the home environment and caregiver practices suggest that the intervention was effective at changing many of those aspects of the child's early life that it sought to influence. Two years into the intervention, homes in the treatment area had a greater diversity of play materials (+0.27 sd), and both mothers and fathers had more frequent and varied interactions with the child (+0.46 sd and +0.33 sd, respectively). Both mothers and fathers increased their use of positive parenting strategies, and decreased their use of punishment behaviours (changes of over 0.3 sd for mother; effect sizes for fathers roughly

half of those for mothers).

An established evidence base, largely from high-income countries, has demonstrated the effectiveness of three early intervention strategies for the promotion of early child development: intensive home visiting, high quality centre-based early childhood education, and nutritional supplements for pregnant women (Currie and Almond (2011)). A number of recent studies have tested programmes with these elements in low- or middle-income settings, including high-quality pre-school programming (Ozler et al. (2018) in Malawi; Chujan and Kilenthong (2021) in Thailand) and home visiting (Attanasio et al. (2014); Andrew et al. (2018) in Colombia; Grantham-McGregor et al. (2020) in India).

In keeping with previous studies, this recent evidence suggests that the details of programme design and implementation are crucial – and if there is a low-cost intervention that is guaranteed to deliver sustained gains in child development, it has yet to be demonstrated. While Chujan and Kilenthong (2021) measure substantial short-term gains from an improved preschool curriculum in Thailand, Ozler et al. (2018) find no effect of training or incentivising teachers at preschools in Malawi. In India, Grantham-McGregor et al. (2020) find that weekly home visiting improved cognitive and language skills of young children. Meanwhile, Attanasio et al. (2014) estimate that a psychosocial stimulation delivered by home visitors in Colombia had substantial effects on cognitive development in the short term; however, these effects had faded two years later (Andrew et al. (2018)).

Despite these mixed findings, caregiver training programmes are emerging as a promising and cost-effective approach to promoting early child development at scale. Using data from the home visiting intervention in Colombia, Attanasio et al. (2020) find that the early successes of the programme were achieved through changes in parents behaviour, rather than from any direct benefit to the child from stimulation by the home visitor. In India, Grantham-McGregor et al. (2020) test the impact of home visiting versus group parenting session, with both treatment arms delivering the same curriculum at the same intensity. The authors find that both interventions had similar intention-to-treat effects, but dramatically different costs: home visiting cost 3.5 times more per household. While Ozler et al. (2018) found no effect of two interventions aimed at pre-primary school teachers, a fourth arm of their study which combined teacher training and group-based sessions for caregivers measured significant improvements in child development 18 months later.

In contrast to much of the existing the literature, this paper focuses solely on a group-

based caregiver training intervention. As shown by Grantham-McGregor et al. (2020), such programmes have considerable potential for scaled implementation, due to their relatively low cost and promising evidence of effectiveness. This paper contributes to this literature by reporting on the short-term effects of just such a programme. Designed on the nurturing care framework, the intervention is positioned to address the gaps in existing scaleable programming identified by leaders in early child development research (see, e.g. Britto et al. (2017)). The programme's replicable model - with a core curriculum adaptable to different local contexts - as well as its established place in the portfolio of a large implementing organisation, make evidence on its effectiveness particularly useful for future implementation.

The remainder of the paper proceeds as follows. Section 2 describes the intervention and the study design. Section 3 describes the data collection, the variables of interest, and balance and attrition. Section 4 describes the empirical framework for the analysis, while Sections 5 and 6 present and discuss the main results and extensions, respectively. Section 7 concludes.

2 Intervention

2.1 Tuwekeze Pamoja

Tuwekeze Pamoja is a pilot study of interventions that combine several of Save the Children's 'Common Approaches' to improving children's learning outcomes. 'Common Approaches' are Save the Children's best understanding of how to address a particular problem facing children. They are based on evidence and can be adapted to work in multiple contexts and also replicated in different countries. Studies such as Tuwekeze Pamoja provide context-specific learning about adaptation, contribute to the existing evidence base, and help inform the further development of these approaches.

Tuwekeze Pamoja is designed to promote child development and school readiness in low-resource, low-academic achievement environments, in order to ensure all children achieve a successful transition to primary school. The suite of interventions supports children from birth through to the first years of primary school. In early childhood, this support comes in the form of community-based caregiver sessions and home visits, with an emphasis on stimulation and nurturing care; as the children grow older, the focus of parenting sessions shifts to transition to school, complemented by teacher professional development for pre-

primary teachers. The approach is designed with scale-up in mind, and as such strives for a model which can be easily replicated, keeping intervention costs per child modest.

The Tanzanian pilot of Tuwekeze Pamoja was launched in 2017 in Mbozi District, Songwe Region, through a partnership between Save the Children and local NGO ADP Mbozi. The project was scheduled to roll out components of the intervention sequentially over a five year period, starting with caregiver training sessions in the first years, and adding teacher and school components in the last two years. This study considers only the first two years of implementation, from baseline data collection in early 2018 to the first follow-up in late 2019. Over this period, caregiver training was the focus of the intervention.

Based on the nurturing care framework, two caregiver training curricula were developed in collaboration with local stakeholders: one targeting caregivers of children aged 0-3, and the other ages 4-6. These curricula were delivered by trained community members through weekly group meetings of 20-25 attendees. Each cycle of sessions for 0-3 year-olds ran for 10 weeks, while the 4-6 sessions ran for 8 weeks. The session content was specific to each age range, but both curricula revolved around responsive caregiving, early learning, nutrition, and child protection. Families identified as being particularly vulnerable were also offered two home visits per cycle.[1]

In addition to this core set of programming, the implementation team developed an additional intervention package designed to reduce violence in the home, both towards children and between spouses. Violence against children is a recognised issue in Tanzania: a 2011 report found that almost three quarters of children under 18 had experienced physical violence, with one quarter having experienced emotional violence (UNICEF (2011)). This 'Plus' package included five additional caregiver sessions focussed on violence-reduction (including conflict resolution, stress management, and gender & parenting) as well as community theatre events, engagement with local leaders, and one further home visit for vulnerable families. It was rolled out as a supplement to the core programming in half the treatment areas.

2.2 Study design

The Tuwekeze Pamoja pilot evaluation adopted a quasi experimental approach: treatment areas were chosen for programmatic reasons, and a set of control areas were then chosen

[1] Families were considered vulnerable if they met any of the following criteria: mother or father less than 17 years old or caregiver over 60 years old; caregiver or child with disability; living in a very remote area.

to match these as closely as possible. For practical reasons, implementation was initially determined at the ward level. These sub-district administrative units are relatively new and somewhat fluid:[2] in the study area they include on average 3-5 villages. Eight wards were chosen for treatment using two criteria: first, the ward must include a health centre; from these, the implementation team selected a range of wards covering the geographic and socioeconomic diversity of the area.[3] These eight wards were further split into Core and Core Plus treatment arms, with the aim of balancing the ward characteristics across arms as much as possible. The eight treatment wards included 35 villages; a further 35 villages were then selected from eight untreated wards with similar characteristics to the treated wards. Figure 1 shows the study location within Tanzania, as well as the Treatment and Control areas within Mbozi district.

3 Data

3.1 Data collection

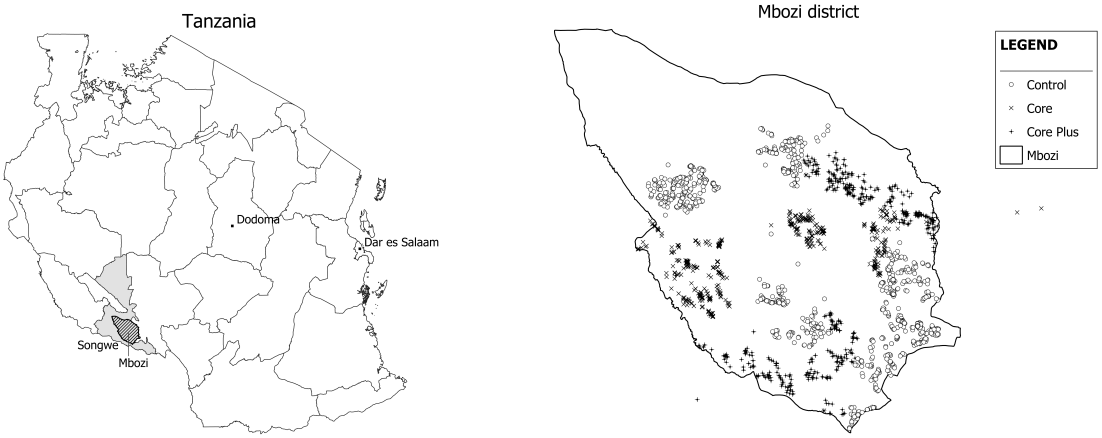
The primary quantitative data collected as part of the pilot are a panel of child-caregiver dyads, to be surveyed every two years over the life of the project. The panel is focussed on a group of children who were infants when the project began. Since the intervention was focussed on caregiver training in the first years, and will only later introduce school-based programme elements, the intervention will 'grow with' the panel, adding elements as the panel children grow up.

The present paper covers only the first two years of intervention and first two waves of data collection: baseline (wave 1), when the panel child was between 4-12 months old;

[2]Songwe region itself was only established in 2016. Wards within the region are reviewed every year based on population and social services, and the boundaries may then be redrawn. While treatment was initially established at the ward level, the village was maintained as the fundamental treatment during implementation. That means that even if ward boundaries adjusted to include or exclude villages from wards initially allocated to treatment or control, all villages maintained treatment status from baseline.

[3]While baseline data allows a comparison of socioeconomic data of respondents across treatment and control, the health centre criteria raises some concerns about the comparability of the areas. No formal record of the criteria used for selection were retained; however, a list of health facilities for all of Tanzania is available from Tanzanian Government Ministry of Health (http://hfrportal.moh.go.tz). Using the August 2021 version of this list, there are 10 health centres and and 45 dispensaries in Mbozi district. Of the health centres in Songwe, two are in treatment wards, three are in control wards, and the five others are in wards outside the study area. Of the dispensaries, nine are in treatment wards and seven are in control wards. This data indicate that five out of eight treatment wards, and four out of eight control wards, have at least one of these two types of health facility. It may be that the the selection of treatment areas at baseline relied on a different definition of health centre; regardless, this data is reassuring regarding the comparability of the treatment and control areas on at least one measure of health care access.

Figure 1: **Study area with Treatment and Control locations**



Maps plotting the study area in the context of Tanzania (L), and Treatment and Control locations within Mbozi district (R). Although Treatment and Control areas were assigned based on ward boundaries at the project inception in 2017, ward boundaries are reviewed annually. Rather than showing ward boundaries, Treatment and Control areas are represented by through GPS data from the baseline survey. These data are somewhat noisy: some points appear incorrectly outside the study area. Map credit: Himangshu Kumar; base maps from https://data.humdata.org.

and the first follow up (wave 2), when the child was between 2-3 years. The first wave of data collection included 2,289 children; 1,721 (75.2%) were re-surveyed at wave 2. Attrition was slightly higher than anticipated, and was due to a mix of temporary absence of the respondent caregiver on the survey day, and permanent absence (due to death or migration) of the child or caregiver.

The panel was initiated at baseline, based on a child of the target age being present in the household. When an eligible household was identified, the enumerator requested to speak with that child's primary caregiver. In the vast majority of cases, the self-reported primary caregiver was the child's mother; in about 5% of cases it was the child's father or, rarely, another caregiver. If the caregiver consented to participate, they were administered two surveys: first, the long form of the Caregiver Reported Child Development Instrument (CREDI, see McCoy et al. (2017, 2018b)), and second, a set of questions about the caregivers' knowledge, activities and practices (KAP). During the follow-up survey, enumerators sought out the same respondent as at baseline, and administered the same two questionnaires.[4]

The long form CREDI questionnaire is a series 108 of yes/no questions about the child, with an age-dependent start point and an endogenous end-point (the survey ends when the

[4]Further details about the data collection can be found in the project Baseline Report (Save the Children (2018)) and Snapshot of Midline Findings (Save the Children (2020)).

caregiver responds negatively to a certain number of questions in a row). The tool is designed to capture motor, cognitive, language and socio-emotional development of children aged 0-3 years old. The tool was extensively tested against a reference group of children raised in 'ideal' home environments, providing a build-in benchmark for normal child development (McCoy et al. (2018a)). Thanks to this, the CREDI tool can be used to generate age-referenced normalised development scores, both as an overall measure and separately for the four domain-specific scores.

The caregiver survey includes a wide range of questions about the child and family, including basic demographics, home environment, activities done with the child, the nature of interactions between child and caregivers, as well as a battery of questions on caregiver attitudes. The survey also includes questions on interaction between partners, caregiver confidence as a parent, as well as Progress out of Poverty's *Poverty Probability Index* (PPI) as a measure socioeconomic status (scoring done based on the 2011 PPI tables for Tanzania (Innovations For Poverty Action (2016)); for a validity assessment of this measure see, e.g. Desiere et al. (2015)). Many of these questions are asked for a range of caregivers (e.g. activities and interactions with mother, father, and other caregiver), while others only concern the respondent caregiver (e.g. attitudes and self-confidence).

3.2 Descriptive statistics

3.2.1 Characteristics of the sample

Table 1 presents some basic demographic characteristics of the panel, as observed at baseline. Just over half the children are girls, with a mean age of 7.7 months. 30% of the children are the respondent caregiver's first child. 95% of respondent caregivers are female. Most of the children's parents are reported to be literate: 84% of mothers, and 92% of fathers. Mother's age was collected using a coarse set of categories: 52% of the caregivers reported the mother to be in the 22-26 year age range or younger.

As described in Section 3.1, the CREDI tool generates age-referenced development scores, normalised against a reference population. Table 2 gives an overview of these scores. At baseline, when the panel children were 4-12 months old, both the treatment and control groups scored on average very close to reference population: the mean CREDI score is approximately zero in both samples. At the second wave, when the children are aged 2-3, stark differences emerge: the control group has now fallen slightly behind the reference

Table 1: **Predetermined characteristics: panel at wave 1**

Variable	Mean	Std. Dev.	N
Girl child	0.534	0.499	1721
Age in months	7.672	2.575	1721
First child	0.295	0.456	1721
Female caregiver	0.949	0.22	1721
Young mother	0.518	0.5	1721
Mother literate	0.84	0.367	1719
Father literate	0.915	0.279	1710

Notes: table shows baseline data from the panel observations.

population, scoring 0.08 standard deviations below average. The treatment group, on the other hand, has pulled ahead by 0.2 standard deviations.

Table 2: **Normalised CREDI scores**

	Treatment	Control	N
Wave 1	0.056	0.032	2,289
	(0.986)	(0.870)	
Wave 2	0.204	-0.084	1,716
	(1.030)	(0.914)	

Notes: values in standard deviations of the CREDI reference population.

3.2.2 Derived variables

The caregiver survey includes a rich set of questions on the home environment and caregiver practices. To facilitate analysis, these data are aggregated into seven index variables: one covering the home environment, and six on caregiver-child interactions and parenting practices (three indices, computed seprately for mother and father).

Providing a stimulating environment for infant children was a particular focus of the 0-3 year old caregiver training sessions; one measure of this environment is the variety of play materials available to the child. Caregivers were asked about nine different categories of play materials: these yes/no answers are summed to an index ranging from 0-9 (a list of the nine categories can be found in Appendix Table 13). This index is then standardised against the control group in each wave.

The respondent was asked about how often, in the past week, the panel child did certain

activities with their mother and their father.[5] Nine activities were listed, including reading books, singing songs, playing games, and taking the child out of the home (see Appendix Table 14 for the full list); most of these also include frequency data (from never, to more than three times in the last week). These data are used to create two indices of parent-child interactions: one for mother and one for father. Each individual activity variable is normalised against the control group in that wave; the nine standardised variables are then averaged to create a single index. This index is then, in turn, standardised against the control group for each wave: the units for these variables are therefore (control group) standard deviations.

A similar approach is used to create indices of parenting style for father and mother. The respondent was asked how often in the last month the mother and the father tried certain things with the child. There are 14 of these questions, eight focussed on positive parenting, and six which ask about disciplinarian actions. Some examples of positive parenting questions are: show affection to your child; explain why a certain behaviour is wrong; listen to what your child thinks. The disciplinarian ('negative') questions include: speak negatively to the child; shake him/her; spank hit or slap your child for misbehaving. (A full list of these questions is found in Appendix Table 15.) As above, the frequency responses to each question are normalised to the control group mean and standard deviation; these standardised variables are then averaged to create indices, which are in turn standardised against the control group for that wave. A total of four indices are created in this way: a separate positive parenting and negative parenting index for mother and father.

3.3 Balance & attrition

3.3.1 Balance

While the treatment and control areas were not selected at random, balance checks at baseline suggest that the resulting sample of individuals is highly comparable on objective measures (a detailed comparison can be found in Appendix A.2). There are some differences in responses to subjective questions on home environment and caregiver practices across groups, with the treatment group more likely to respond affirmatively to all questions. This pattern of differences suggests that there may be an underlying difference in the propensity

[5] These questions, as well as the ones on parenting style, were also asked for "other" caregiver, if relevant; responses to this were often missing, likely due to the child not having another primary caregiver. These data are excluded from analysis in this paper.

to respond, rather than a difference in the respondents themselves.

Two account for these differences, the primary specification in all regressions controls for baseline levels of both demographic as well as home environment and caregiver practice variables. Furthermore, to explore the potential impact of this imbalance, the main results are estimated using two sets of baseline controls: demographics only, and the full set of controls. There is little difference in the results between these two models, suggesting that these baseline differences are not critical to the results.

3.3.2 Attrition

Of the original sample of 2,289 children, 1,721 (75.2%) were re-interviewed during the follow up. Attrition rates were very similar across treatment and control (24.5% vs 25.1%). The attrited observations are somewhat different from panel members: they are more likely to be first children and to be children of young mothers, and less likely to be girls or children of literate mothers (a detailed comparison can be found in Appendix A.2, Table 17). There are no statistically significant differences between attrited and panel observations on baseline child development scores, or on the home environment and caregiver practice variables.

There are some differences in attriters across treatment and control. Attriters in the treatment group are slightly older, more likely to have a young mother and be first born. In contrast they are less likely to have literate parents, and come from lower SES households (see Appendix A.2, Table 18). This suggest that attrition in the control group is more positively selected, with older, more established, more educated and wealthier households dropping out, as compared with attriters in the treatment group.

While the inclusion of a rich set of control variables will help address this, the robustness of the results to selective attrition is explored in two ways. First, inverse probability weights are applied to restore the original characteristics of the sample. Second, Lee bounds (Lee (2009)) are estimated on a simplified version of the model. Very similar treatment effects are estimated in each case.

4 Empirical framework

4.1 Estimation strategy

This paper seeks to estimate the impact of the Tuwekeze Pamoja intervention. The intervention was designed to promote early child development: the primary outcome of interest is therefore the summary measure of child development from the CREDI questionnaire. We also explore two further empirical questions: first, did the additional 'Plus' package of interventions, designed to reduce violence, have any effect on child development beyond the Core intervention? And second: through what channels did the intervention achieve change in child development?

The estimation strategy adopted here relies on the assumption that, conditional on observables, treatment was assigned as good as randomly. The balance of demographic covariates at baseline suggest that this assumption is reasonable: no substantial differences are found on these variables. The balance on home environment and caregiver practices raises some concerns, as the baseline levels of these are higher in the treatment group than in the control group. To identify the causal effect of the intervention on child development, the estimation strategy will need to be able to fully control for any independent effect these baseline differences might have on on the outcome variable. The panel nature of the data facilitates the inclusion of a generous set of control variables; however, identification relies on these controls capturing all relevant differences between treatment and control.

4.2 Estimating equations

4.2.1 Primary specification: intention to treat

The primary estimating equation, which estimates the intention to treat effect of the intervention by comparing treatment and control areas, is:

$$Y_{i2} = \alpha_0 + \alpha_1 \text{treat}_i + \alpha_2 Y_{i1} + \delta X_{it} + \epsilon_{it}, \tag{1}$$

where Y_{it} is an individual outcome of interest measured at waves $t = 1, 2$; $treat$ is a binary variable; and X_{it} is a set of predetermined control variables. When estimating the primary treatment effect of the intervention, Y_{it} is the overall child development score; when investigating the channels through which the intervention was effective, Y_{it} is a caregiver-level

outcome. In both cases, the coefficient of interest is $\hat{\alpha_1}$, the estimate of α_1.

To estimate whether the violence-reduction package had any additional effect, Equation 1 is modified to estimate the treatment effect in the two arms separately, as follows:

$$Y_{i2} = \beta_0 + \beta_1 \text{Core}_i + \beta_2 \text{Core Plus}_i + \beta_3 Y_{i1} + \gamma X_{it} + \nu_{it}. \quad (2)$$

In Equation 2, the coefficients of interest are the estimates of β_1 and β_2.

4.2.2 Extension: treatment effect on the treated

When the second wave of data was collected, not all caregivers in the treatment area had attended the training sessions.[6] This was due to two things: first, incomplete roll-out of the intervention and second, caregivers choosing not to attend the sessions that were offered in their local area. While the intervention aims to offer both the 0-3 and 4-6 caregiver training sessions on an annual basis in each hamlet of the treatment area,[7] these sessions were rolled out across locations based on the availability of the trained community facilitators delivering the sessions.

The primary focus of this analysis is the Intention to Treat (ITT) estimator: in general, this is the best estimate of the impact the intervention would have if it was rolled out at scale. This may not be the case, however, if incomplete compliance is due largely to incomplete roll-out of the intervention in the treatment area. If all caregivers attend sessions when they are offered (imperfect compliance is due solely to incomplete roll-out), then the estimated impact of the treatment on the treated (ToT estimate) is a better estimate of the impact intervention would have when it is running as planned, e.g. the ITT if roll-out had been complete. If, on the other hand, only some caregivers attend, then even when roll-out is complete there will only be partial compliance, pushing the ToT estimates above the ITT. Furthermore, compliance is likely to be selective: those who choose to attend will not be a random sample of all caregivers, and the treatment effect on that group is likely to differ from the mean.[8] If the effect of the intervention is heterogenous this would drive a further positive wedge between the ITT and ToT estimates.

[6] There is no evidence of any caregivers in control areas ever attending sessions. While this is theoretically possible, in practice the sessions drew from a close geographic area.

[7] Villages in the area typically have 3-6 hamlets. In some cases sessions were offered to residents of a single hamlet, while in other cases they drew from more than one hamlet.

[8] In this particular case, there is little evidence that those who attended sessions differed substantially from those who did not on observable characteristics (see Appendix Table 20); however, they may also have unobserved characteristics which make them particularly receptive to the intervention itself.

With these concerns in mind, we also estimate the treatment effect on the treated. During the second wave, respondents were asked whether one of the child's caregivers attended any of the training sessions. We estimate the impact of the intervention on those who attended using two-stage lease squares, where the endogenous choice to attend sessions is instrumented by the treatment status. The estimating equations are given by:

$$D_i = \beta_0 + \beta_1 \text{treat}_i + \gamma X_{it} + \nu_{it} \tag{3}$$

$$Y_{i2} = \alpha_0 + \alpha_{2SLS}\hat{D}_i + \alpha_2 Y_{i1} + \delta X_{it} + \epsilon_{it} \tag{4}$$

where D_i is attendance at one or more caregiver training sessions and $treat_i$, treatment status, is the instrument excluded in Equation 4. The other variables are as defined in Equation 1. The coefficient of interest is $\alpha_2\hat{s}_{LS}$, the estimated treatment effect on those who took the treatment.

4.2.3 Accounting for multiple hypotheses

While a single variable summarises the primary outcome of interest – early child development scores – the exploration of any further outcome variables requires the testing of multiple hypotheses at the same time. This is the case for two sets of outcomes: first, when looking at the subdomains of child development individually; and second when considering the range of indicators of home environment and caregiver practices. As standard tests of statistical significance are designed with a single hypothesis test in mind, testing multiple hypothesis using these same tools can increase the probability of falsely rejecting any one null hypothesis: specifically in this case, falsely finding a treatment effect to be statistically significantly different from zero.

To correct for this, we apply the Romano-Wolf multiple-hypothesis correction, as introduced by Romano and Wolf (2005b,a), and implemented in Stata by Clarke et al. (2020) and Clarke (2021). The Romano-Wolf correction accounts for the fact that a set of hypothesis tests are related, and seeks to control the family-wise error rate: the probability of falsely rejected at least one true null hypothesis amongst this set. This correction method uses re-sampling to estimate the correlation between test statistics, allowing for a less conservative adjustment of p-values than correction methods that assume these to independent such as

the Bonferroni correction (Clarke et al. (2020)). Results tables for which the Romano-Wolf multiple-hypothesis correction is applied show both naive (model) p-values and corrected ones, with details of the correction parameters in the tables notes.

5 Main results

5.1 Child development

The intervention was designed to promote early child development: the primary outcome under consideration is therefore the overall child development score (CREDI). Table 3 presents estimates of Equations 1 and 2 with child development scores as the dependent variable: first combining the two treatment arms (Column (1)), and second separately estimating the impact of the Core and Core Plus treatments (Column (2)).

The overall treatment effect combining both treatment arms is 0.255, and highly significant (Table 3, Column (1)). Recalling that CREDI measures child development in standard deviations of a reference population, this indicates that children in the treatment group had 25.5% of a standard deviation higher child development scores, compared with similar children in the control group.

There is no statistically significant difference in the treatment effect across the two treatment arms (Table 3, Column (2)), which have almost identical point estimates (0.248 vs 0.262). This shows that the additional components of the Core Plus intervention did not significantly improve child development outcomes, with respect to the Core intervention package. Given this finding, the remainder of the analysis will focus on a comparison of treatment (combining both arms) and control.

The models estimated in Table 3 also gives insights into the baseline covariates which are associated with child development two years later. Considering the primary specification in Column (1), we find that child development scores are persistent, although weakly so: a 1 standard deviation increase in CREDI at baseline is associated with 0.168 standard deviation higher CREDI scores 2 years later. This is similar in magnitude to the coefficient on mother's literacy (0.165), but considerably more than that on socio-economic status (SES): a 1 standard deviation increase in SES is associated with a 0.064 increase in CREDI scores. The indicator for having a young mother has the largest magnitude coefficient, associated with a -0.194 difference in chid development scores, with respect to having an

older mother.

A number of the baseline home environment and parenting practice variables show a positive association with child development scores. These variables are all positively correlated (see Appendix Table 19), although far from perfectly so. Index variables for the home environment, mother-child interactions, and father's use of positive parenting practices are all positively associated with child development; the other variables, including both mothers' and fathers' use of negative parenting practices, have no statistically significant association.

Table 3: **Child development: normed CREDI scores**

	(1) Overall		(2) Overall	
Treat	0.255***	(0.0571)		
Core			0.248***	(0.0613)
Core Plus			0.262***	(0.0620)
Overall baseline	0.168***	(0.0335)	0.168***	(0.0336)
First child	0.0696	(0.0479)	0.0694	(0.0477)
Female caregiver	-0.0968	(0.109)	-0.0974	(0.108)
Girl child	0.0351	(0.0447)	0.0349	(0.0447)
Age in months	0.0152	(0.0107)	0.0152	(0.0107)
SES (in sd)	0.0637**	(0.0264)	0.0635**	(0.0265)
Young mother	-0.194***	(0.0616)	-0.193***	(0.0613)
Mother literate	0.165*	(0.0824)	0.164*	(0.0832)
Father literate	0.0926	(0.0903)	0.0930	(0.0900)
Kinds of toys	0.0643**	(0.0290)	0.0646**	(0.0292)
Mother Interaction	0.0714***	(0.0201)	0.0713***	(0.0203)
Father Interaction	-0.0292	(0.0225)	-0.0293	(0.0224)
Mother PP	-0.0235	(0.0316)	-0.0235	(0.0316)
Mother NP	0.0238	(0.0278)	0.0236	(0.0276)
Father PP	0.0526*	(0.0291)	0.0524*	(0.0289)
Father NP	0.00128	(0.0224)	0.00135	(0.0223)
Constant	-0.277	(0.175)	-0.277	(0.175)
Observations	1672		1672	
F-T1vsT2			0.0922	
F-pval			0.766	

Notes: dependent variable is in standard deviations of reference population. Standard errors in parentheses; * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$.

In addition to providing an overall measure of child development, the CREDI score can be broken down into four component domains capturing motor, cognitive, language and socio-emotional development. Table 4 presents the treatment effects from estimates of Equation 1, with the overall score and each domain score as the dependent variable. Given that

these subdomains were not the primary target of the intervention, to account for multiple-hypothesis testing Table 4 reports both the model p-values, and Romano-Wolf corrected p-values (this correction was done simultaneously for the treatment effects estimated from the full set of 12 models reported across Tables 4 & 5).[9]

The estimated treatment effects on child development across the four domains are all quite similar, ranging from 0.21 (socioemotional) to 0.27 (language). It is not clear how sensitive these domain-level measures are, so the results should be interpreted with caution. They indicate, however, that the intervention was effective at improving child development across this full range of domains. The larger coefficients on cognitive and language development are suggestive of a particular impact on those areas.

Table 4: **Child development subdomains: normed CREDI scores**

	(1) Overall	(2) Motor	(3) Cognitive	(4) Language	(5) Socioemotional
Treat	0.255***	0.218***	0.257***	0.273***	0.210**
	(0.000)	(0.010)	(0.007)	(0.000)	(0.022)
	[0.0001]	[0.0007]	[0.0006]	[0.0001]	[0.0012]
Controls	Yes	Yes	Yes	Yes	Yes
Observations	1672	1672	1672	1672	1672

Notes: dependent variable is in standard deviations of reference population. Model p-values in parentheses; Romano-Wolf p-values in square brackets (testing all hypotheses shown in Tables 4 & 5 together; 10000 bootstrap replications). Statistical significance indicated based on Romano-Wolf: * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$.

5.2 Home environment and caregiver practices

The previous section demonstrated that the intervention led to substantial improvements in early child development. Given that the intervention was designed to achieve these gains through changes in the home environment and child-caregiver interactions, we would expect to see treatment effects on these variables as well. To explore these channels, we re-estimate Equation 1 with the targeted home environment and caregiver practice indices as the dependent variables. Recall that these indices are measured in standard deviation units of the control group in each wave.

Table 5 shows the estimated treatment effects on these seven indicators. All show statis-

[9] Given the close similarity between the model (uncorrected) and Romano-Wolf p-values in Tables 4 & 5, extensions and further results in Section 6 are reported along with conventional standard errors, with statistical significance assessed using model p-values.

tically significant changes in the expected direction: improvements in the diversity of play materials in the home, increase in interactions of mother and father with the child, increases in positive parenting practices and decreases in negative parenting. The magnitude of these treatment effects range from 0.46 to 0.18 standard deviations, with larger changes seen for mothers than for fathers.

These treatment effects are supportive of the intervention's theory of change: caregiver training can effectively improve child development by changing the child's home environment and family interactions. The larger behaviour changes for mothers is also consistent with the design of the programme, as mothers were far more likely to attend caregiver training sessions than fathers. It is important to keep in mind, however, that all seven of these variables are self-reported, and had higher baseline levels in the treatment group. The design of data collection does not allow us to estimate the degree to which these changes could be due to social desirability or other biases arising from the engagement the treatment group has had with the intervention and with early child development concepts. While the CREDI is also caregiver-reported, the nature of the questions (which ask about specific things the child can or cannot do) make them less sensitive to this concern.

Table 5: **Home environment and caregiver practices**

	(1) Toys	(2) Mother Int	(3) Father Int	(4) Mother PP	(5) Mother NP	(6) Father PP	(7) Father NP
Treat	0.265***	0.457***	0.333***	0.357***	-0.318***	0.187**	-0.175**
	(0.0003)	(0.0000)	(0.0002)	(0.0057)	(0.0010)	(0.0180)	(0.0268)
	[0.0001]	[0.0001]	[0.0001]	[0.0006]	[0.0001]	[0.0012]	[0.0012]
Controls	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Observations	1682	1688	1688	1643	1640	1566	1562

Notes: dependent variable is in standard deviations. Model p-values in parentheses; Romano-Wolf p-values in square brackets (testing all shown hypotheses in Tables 4 & 5 together; 10000 bootstrap replications). Statistical significance indicated based on Romano-Wolf: * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$.

6 Extensions and discussion

6.1 Extensions

6.1.1 Treatment on the treated estimates

During the second wave of data collection, respondent caregivers were asked whether the primary caregiver, or any other of the child's caregivers, attended any of the training sessions.

In the treatment group, 431 (49.4%) of respondents reported that one of the child's caregivers attended the sessions (of these, 406 reported that the primary caregiver attended, and 38 reported that a different caregiver attended - 13 of these indicated that both primary plus another caregiver attended). None of the control group reported attending any sessions.

To estimate the treatment effect of the intervention on those who attended at least one session, the instrumental variables model described in Equations 3 & 4 is estimated, where the endogenous choice to attend is instrumented by treatment status. Table 6 shows the results from an OLS regression of attendance on treatment status and controls. Treatment status is highly predictive of attendance; however, none of the other baseline covariates is. This supports anecdotal evidence from the field that the primary reason for low attendance was incomplete programme roll out: many caregivers had not yet been given the opportunity to attend.

Table 6: **Predicting attendance**

	(1) Attended	
Treat	0.500***	(0.0812)
First child	-0.0460	(0.0312)
Female caregiver	0.0353	(0.0285)
Girl child	0.0102	(0.0198)
Age in months	0.00711	(0.00448)
SES (in sd)	-0.0172	(0.0144)
Young mother	0.0445	(0.0290)
Mother literate	0.0192	(0.0365)
Father literate	0.0192	(0.0358)
Kinds of toys	0.0119	(0.00738)
Mother Interaction	-0.000808	(0.00948)
Father Interaction	0.00109	(0.0119)
Mother PP	-0.0192	(0.0157)
Mother NP	-0.0171	(0.0110)
Father PP	0.00625	(0.0115)
Father NP	0.00276	(0.00810)
Constant	-0.136*	(0.0699)
Observations	1688	
R^2	0.342	

Notes: regression of attendance (binary) on treatment status and other controls. Standard errors in parentheses; statistical significance is indicated based on model p-values. * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Table 7 presents results from instrumental variables estimates of the primary outcome

variables. As expected, the estimated treatment effects are substantially higher under the IV specification, with treatment effects roughly twice as large as the ITT estimates reported in Tables 3 & 5: the estimated effect of the intervention on child development rises from 0.255 sd (OLS) to 0.505 sd (IV).

Table 7: **Summary results: Instrumental variables estimates**

	(1) CREDI	(2) Toys	(3) Mother Int	(4) Father Int	(5) Mother PP	(6) Mother NP	(7) Father PP	(8) Father NP
Attended	0.505***	0.530***	0.915***	0.667***	0.709**	-0.631***	0.366**	-0.343**
	(0.133)	(0.140)	(0.199)	(0.175)	(0.312)	(0.193)	(0.172)	(0.154)
Controls	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Observations	1672	1682	1688	1688	1643	1640	1566	1562

Notes: treatment effect estimates from 2SLS regressions where attendance is instrumented by treatment status. All regressions include controls for basic demographics as well as home environment and caregiver practices. Standard errors in parentheses; statistical significance is indicated based on model p-values. * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

6.1.2 Heterogeneous effects

Do all participants respond to the intervention in a similar way? Although the study was not powered to rigorously estimate the impact of the intervention on subgroups, this remains an important question from a policy perspective. To explore possible heterogeneous treatment effects, we re-estimate the primary specification, including a sequential set of interaction terms between treatment status and predetermined child or household characteristics.

Table 8 reports the main and interaction effects for each of these regressions. Perhaps surprisingly, none of the interaction effects is statistically significant, suggesting that any heterogeneity in treatment effects is relatively small. The two interaction terms with the largest magnitude coefficients are the indicator for female respondent, and the indicator for literate fathers. The point estimate for female respondents is large and negative, suggesting that those households in which a male self-reported as primary caregiver showed larger treatment effects; the point estimate for literate fathers is large and positive, also suggesting an increase in treatment effect. Both are imprecisely estimated and not statistically different from zero, but could be interesting sources of heterogeneity to investigate in future.

We carry out a similar exercise for the baseline home and caregiver characteristics. The imbalance at baseline in these variables across treatment and control suggests a differing disposition to report this information across groups. We argue that the child development questions, being more objective, are less vulnerable to response biases; however it could be that those caregivers who report higher values on the home environment and caregiver

Table 8: **Child development: heterogeneity by demographics**

	(1) CREDI	(2) CREDI	(3) CREDI	(4) CREDI	(5) CREDI	(6) CREDI	(7) CREDI	(8) CREDI
Treat=1	0.276***	0.381	0.280**	0.278	0.254***	0.242**	0.269	0.108
	(0.0572)	(0.264)	(0.0841)	(0.136)	(0.0569)	(0.0660)	(0.153)	(0.156)
Treat=1 × First child	-0.0737							
	(0.102)							
Treat=1 × Female caregiver		-0.133						
		(0.263)						
Treat=1 × Girl child			-0.0466					
			(0.0850)					
Treat=1 × Age in months				-0.00300				
				(0.0164)				
Treat=1 × SES (in sd)					-0.0442			
					(0.0498)			
Treat=1 × Young mother						0.0242		
						(0.127)		
Treat=1 × Mother literate							-0.0167	
							(0.139)	
Treat=1 × Father literate								0.160
								(0.153)
Controls	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Observations	1672	1672	1672	1672	1672	1672	1672	1672

Notes: regressions include controls for basic demographics as well as home environment and caregiver practices. Standard errors in parentheses; statistical significance is indicated based on model p-values. * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

indices show a similar bias when answering the CREDI questions. If this were the case, we would expect the higher values of home and caregiver indices in the treatment group to be associated with systematically higher child development scores: if baseline controls are do not net out this difference, it would lead to a positive interaction effect between treatment and home and caregiver indices.

To investigate this, we interact each of the home environment and caregiver indices with treatment status. Treatment effects and interaction terms for these regressions are shown in Table 9. As in the heterogeniety analysis above, none of the interaction terms is statistically significant - and here they are also all small in magnitude. This provides some further reassurance that baseline differences in these variables are not driving the treatment effects.

6.2 Robustness

6.2.1 Balance

While balance on demographic characteristics was quite good at baseline, the imbalance on pre-treatment values of the caregiver practice variables raises some concerns. While the primary specification in all regressions controls for the baseline levels of both demographic

Table 9: **Child development: heterogeneity by targeted practices**

	(1) CREDI	(2) CREDI	(3) CREDI	(4) CREDI	(5) CREDI	(6) CREDI	(7) CREDI
Treat=1	0.253***	0.253***	0.255***	0.256***	0.253***	0.255***	0.252***
	(0.0574)	(0.0585)	(0.0569)	(0.0576)	(0.0568)	(0.0572)	(0.0549)
Treat=1 × Kinds of toys	0.0356						
	(0.0534)						
Treat=1 × Mother Int		0.0906					
		(0.0513)					
Treat=1 × Father Int			-0.0239				
			(0.0613)				
Treat=1 × Mother PP				-0.0253			
				(0.0590)			
Treat=1 × Mother NP					0.0938		
					(0.0744)		
Treat=1 × Father PP						0.0126	
						(0.0488)	
Treat=1 × Father NP							0.0700
							(0.0837)
Controls	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Observations	1672	1672	1672	1672	1672	1672	1672

Notes: regressions include controls for basic demographics as well as home environment and caregiver practices. Standard errors in parentheses; statistical significance is indicated based on model p-values. * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

as well as home environment and caregiver practice variables to account for this, it is informative to explore to what extent this affects the results. To do so, we estimate the main results using two sets of baseline controls: demographics only, and the full set of controls. If these two models give similar results, this suggests that the imbalance in parenting style at baseline is not driving the findings.

Table 10 reports the treatment effect estimated from Equation 1 using three different sets of controls. First, in Column (1), only demographic controls are included while Column (2) reports results from the main specification with both demographic and home & caregiver controls. The treatment effects estimated across these two specification are almost identical, although the point estimate is slightly higher when fewer controls are included: controlling for these baseline differences does not substantially change the estimated treatment effect.

The similarity of the estimates in Columns (1) and (2) might suggest that the intervention's theory of change is weak: if there were baseline differences in home & caregiver variables favouring the treatment group, and if indeed these are instrumental to child development, we would expect that controlling for them would reduce size of the treatment effect. Do home environment and caregiver practices actually matter? To explore this, we

re-estimate the main treatment effect, controlling for wave 2 values of the home environment and caregiver practices: the characteristics the intervention sought to modify directly. The results are reported in Column (3): once the wave 2 values of these variables have been controlled for, the treatment effect is statistically insignificant, and small in magnitude. This provides supporting evidence that it is indeed treatment-induced changes in these variables which are driving the impact of the intervention on child development. It also suggests the baseline differences in these variables between treatment and control groups may not be capturing fundamental differences, but rather a greater propensity to reply affirmatively to questions. Controlling for these additional baseline variables appears to be a sensible strategy, resulting in a slightly more conservative but qualitatively identical point estimate, compared with Column (1).

Table 10: **Child development: varying sets of controls**

	(1) Overall	(2) Overall	(3) Overall
Treat	0.281***	0.255***	0.0856
	(0.0560)	(0.0571)	(0.0633)
Basic Controls	Yes	Yes	Yes
Baseline Targeted Practices	No	Yes	Yes
Endline Targeted Practices	No	No	Yes
Observations	1693	1672	1532

Notes: regressions include varying sets of controls. Standard errors in parentheses; statistical significance is indicated based on model p-values. * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

6.2.2 Attrition

The overview of attrition in Section 3.3 highlighted some differences between panel observations and attriters, as well as some differences between attriters in treatment and control. The characteristics of attriters across treatment and control suggest positive selection into attrition in the control group compared with the treatment group. This pattern of attrition could create a positive bias in the estimated treatment effect, if children who would have had higher development scores are more likely to drop out of the sample if they are in the control group. We explore the sensitivity of the results to bias from attrition in two ways: first, by re-weighting the sample; and second through a bounding exercise.

Given the rich data available at baseline, it is plausible that selection into attrition

is based largely on observable characteristics. If such selection is entirely on observables, then re-weighting the panel observations to restore characteristics of the original sample will address the bias caused by individuals with different characteristics attriting from the panel. To check the sensitivity of the results to this type of correction, we re-estimate Equation 1 with inverse probability weights. The probability of remaining in the panel is estimated from a logistic regression on the baseline data, of the form:

$$panel_i = \lambda_0 + \lambda_1 X_{i1} + \epsilon_i, \tag{5}$$

where $panel_i$ is equal to 1 for those subjects who are observed both waves, and equal to 0 for those who are only in wave 1 (attriters). The X_{i1} are individual characteristics measured at baseline, and include both basic demographics and the baseline values of home environment and caregiver practices variables. Coefficients estimated from Equation 5 are used to predict the probability (ρ) of remaining in the panel for all respondents: the inverse of this probability ($1/\rho$) is then used to re-weight the panel. This approach gives more weight to those panel observations that were least likely to remain in the sample (most likely to attrit).

Table 11 reports the main results, estimated with inverse probability weights. For all estimates, the point values are almost identical to the primary specification. This suggests that any differences in attrition based on observable characteristics are not significantly biasing the estimated treatment effects.

Table 11: **Summary findings: inverse probability weighting**

	(1) CREDI	(2) Toys	(3) Mother Int	(4) Father Int	(5) Mother PP	(6) Mother NP	(7) Father PP	(8) Father NP
Treat	0.253***	0.266***	0.457***	0.331***	0.366***	-0.311***	0.195**	-0.167**
	(0.0595)	(0.0569)	(0.0722)	(0.0708)	(0.109)	(0.0778)	(0.0706)	(0.0715)
Controls	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Observations	1672	1682	1688	1688	1643	1640	1566	1562

Notes: regressions include controls for basic demographics as well as home environment and caregiver practices. Regressions are weighted using inverse-probability weights. Standard errors in parentheses; statistical significance is indicated based on model p-values. * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

As an additional exercise, we estimate Lee Bounds on the main treatment effect, as proposed by Lee (2009) and implemented in Stata by Tauchmann (2018). This procedure re-estimates the treatment effect under two extreme scenarios of unbalanced selection into the panel: a "best case" for the treatment effect (e.g. positive attrition from the control

group, or negative attrition from the treatment group, causing a positive difference in outcomes), and a symmetric "worst case." A limitation is that the procedure is designed for randomised controlled trials, and can only accommodate limited categorical control variables. For illustrative purposes we re-estimate a naive treatment effect using a simplified version of Equation 1, and calculate bounds for that estimator. This gives some sense of the scope for attrition to bias the results.

Table 12 reports treatment effects estimates from this simplified model, along with the Lee bounds around this estimate. With a few exceptions (notably, the estimated effect of the treatment on negative parenting), the Lee bounds for each estimate are all quite close to the point estimate itself. This is likely due in part to attrition across both treatment arms being relatively similar: in each Lee bound estimation, less than 1% of observations were trimmed to balance the two arms.

Table 12: **Summary findings: Lee bounds**

	(1) CREDI	(2) Toys	(3) Mother Int	(4) Father Int	(5) Mother PP	(6) Mother NP	(7) Father PP	(8) Father NP
Treat	0.289***	0.283***	0.470***	0.303***	0.365***	-0.342***	0.173***	-0.208***
	(0.0472)	(0.0491)	(0.0538)	(0.0554)	(0.0550)	(0.0499)	(0.0549)	(0.0494)
Constant	-0.0857**	-0.0000335	0.00150	-0.00319	0.000640	0.000463	-0.00145	0.000236
	(0.0335)	(0.0349)	(0.0383)	(0.0394)	(0.0391)	(0.0354)	(0.0387)	(0.0347)
Observations	1705	1715	1721	1721	1672	1669	1584	1582
Lower bound	0.271	0.283	0.422	0.274	0.350	-0.356	0.164	-0.358
Upper bound	0.311	0.293	0.470	0.312	0.386	-0.187	0.177	-0.192

Notes: treatment effects from a simple regression with no controls.

While neither of these examples can prove that attrition is not biasing the treatment effects estimated on the panel of observations, they provide reassurance that the point estimates of interest - and that for child development in particular - are robust to two different commonly-used approaches to addressing attrition.

7 Conclusion

This paper evaluates the effect of a group-based caregiver-training intervention on early child development outcomes. After two years of implementation, 2-3 year olds in the treatment area scored on average 0.26 standard deviations higher on a holistic child development measure, compared with similar children in the comparison area. At the time of data collection, only half of caregivers in the treatment group had attended any of the training sessions: estimates of the effect of the intervention on children whose caregivers attended

sessions is roughly twice as large as the intention-to-treat estimate.

Although, as a pilot study, the average cost of the intervention was quite high (average cost per child reached is in the range of £115-225), the marginal cost of extending the intervention is much lower (in the range of £9-42). The estimated effect sizes imply a marginal cost effectiveness of between £36-£83 per standard deviation increase in child development (details of these calculations, and cost estimates, are in Appendix B).

Detailed data on caregiver-child interactions and the home environment indicate that the measured changes in child development were driven by caregivers adopting nurturing care practices and providing additional child stimulation. Two years into the intervention, households in the treated area had a greater diversity of play materials, and parents had changed the quantity and quality of their interactions with the young child. These changes were more pronounced for mothers, but were statistically significant and modest in size for fathers as well.

While the settings and intensity of the programmes are different, these results are quite comparable to the improvements in child development found by Grantham-McGregor et al. (2020) from group-based caregiver training in India. Their study found treatment effects of 0.28 sd (cognition) and 0.30 sd (language), at a cost of $38 per child per year over two years (implying an average cost of approximately $271 per sd improvement in cognitive skills). Furthermore, while the programme in India was more intense, with weekly meetings over two years, the authors found that most of the gains were realised in the first year. It is therefore perhaps not surprising that a shorter intervention is able to achieve comparable gains.

This study has a number of limitations. First, the follow up survey was done while the intervention was still underway. This prevents any analysis of the persistence of the effects over time. Two recent studies with medium-term follow-ups found that early gains had dissipated within a few years of the intervention (Ozler et al. (2018)'s study in Malawi found that the most promising arm of pre-school teacher training combined with caregiver sessions had no effects after 36 months, while the initial improvements found in Andrew et al. (2018)'s home visiting programme in Columbia had faded after 2 years).

The timing of the follow-up survey in this study leads to two particular limitations: first, the value of the developmental gains measured during the first two years of the programme depends critically on how durable they are. Second, the Tuwekeze Pamoja intervention is

deliberately designed to sustain these early gains, by supporting children and caregivers throughout early childhood: the curriculum includes training for caregivers of children aged 0-3 and 4-6, with further programming for teachers when these children reach primary school. This feature of the intervention has the potential to address the fade-out that has plagued previous interventions. By focussing here on the first two years, and following a cohort of children from ages 0 to 3, the analysis here gives only a partial picture of the impact of the intervention when fully implemented.

A related limitation is that the analysis here cannot estimate the impact of elements of the intervention that did not engage caregivers of children aged 0-3. Specifically, it is not possible to evaluate the impact of the caregiver training sessions targeting children aged 4-6, nor to comment on the relative efficacy of intervening at different ages. While there is substantial evidence that the 0-3 age range is particularly critical (see, e.g. Britto et al. (2017)), from a programme perspective it would be valuable to know the return to each of these segments of the intervention, both individually and when combined.

Finally, this study is limited by its reliance on self-reported data. This is a particular concern when analysing data from the caregiver survey, which included questions which are highly subjective and therefore vulnerable to a range of reporting biases. In addition to the issues of imbalance discussed in the paper, responses to these questions could be particularly affected by social desirability bias on the part of respondents in the treatment group, whose views on the socially appropriate responses to these questions could be shaped by the intervention itself. Fortunately, the primary outcome variable - the CREDI measure of child development - is collected through a series of questions of a more objective nature; however, these are still reported by the primary caregiver.

The results reported in this paper have two important policy implications. The headline findings demonstrate the potential for caregiver-focussed interventions to have a substantial impact on early child development. The potential of such interventions is significant: not only do they provide an opportunity to reach children from the earliest ages, including the critical 0-3 period; they also provide a practical policy option in areas where centre-based early childcare is not widely available.

Furthermore, data from caregivers suggests that the nature of relationship between caregivers and their children is highly malleable. The Tuwekeze Pamoja intervention was not especially intensive, and yet resulted in changes in caregiver practices across a wide range

of measures. It is likely that the details of the intervention are quite critical here - in particular, the process of adapting the core curriculum to the local context. It would therefore be wrong to extrapolate these findings to caregiver training programmes in general; nevertheless, these results suggest that caregivers are perhaps more receptive to change than previously thought.

The promising short-term results reported in this paper raise a number of further research questions. The most directly relevant to Tuwekeze Pamoja, and for similar programming under consideration at Save the Children, is the question of persistence. Do these early gains translate into improved school-readiness at ages 5-6 and, eventually, to a successful transition into primary school? As a driving motivation for the intervention, answers to these questions are critical for assessing the long-term value of the programme.

Second, while the analysis here has highlighted a number of areas in which caregivers changed their behaviour in response to the intervention, the study was not designed to identify which of these changes had the greatest impact on child development - nor which aspects of the intervention were most responsible for generating change. Further pilot studies should seek to shed light into these two black boxes, as this information would help future programming maximise effectivenss, and ensure that adaptations of the intervention to new context preserve key elements.

Finally, as with any single-site pilot study, important questions remain about whether similar programmes can achieve similar gains in other contexts (Sabol et al. (2022)). Rural Tanzania shares many features with low-resource, low-primary school achievement areas around the world; however, further research is needed to understand in which type of settings a programme like Tuwekeze Pamoja will replicate these successes. Given that the programme relies on the openness of caregivers to adopting new parenting approaches, variations in this across cultures could be a critical feature to consider.

References

Andrew, A., O. Attanasio, R. Bernal, L. C. Sosa, S. Krutikova, and M. Rubio-Codina (2019): "Preschool Quality and Child Development," 78.

Andrew, A., O. Attanasio, E. Fitzsimons, S. Grantham-McGregor, C. Meghir, and M. Rubio-Codina (2018): "Impacts 2 years after a scalable early childhood development intervention to increase psychosocial stimulation in the home: A follow-up of a cluster randomised controlled trial in Colombia," *PLOS Medicine*, 15, e1002556 – 19.

Attanasio, O., S. Cattan, E. Fitzsimons, C. Meghir, and M. Rubio-Codina (2020): "Estimating the Production Function for Human Capital: Results from a Randomized Controlled Trial in Colombia," *American Economic Review*, 110, 48 – 85.

Attanasio, O. P., C. FernÃ¡ndez, E. O. A. Fitzsimons, S. M. Grantham-McGregor, C. Meghir, and M. Rubio-Codina (2014): "Using the infrastructure of a conditional cash transfer program to deliver a scalable integrated early child development program in Colombia: cluster randomized controlled trial," *BMJ*, 349, g5785 – g5785.

Black, M. M., S. P. Walker, L. C. H. Fernald, and e. al. (2017): "Early childhood development coming of age: science through the life course," *The Lancet*, 389, 77 – 90.

Britto, P. R., S. J. Lye, K. Proulx, A. K. Yousafzai, and e. al. (2017): "Nurturing care: promoting early childhood development," *The Lancet*, 389, 91 – 102.

Chujan, W. and W. T. Kilenthong (2021): "Short-Term Impact of an Early Childhood Education Intervention in Rural Thailand," *Journal of Human Capital*, 15, 269–290.

Clarke, D. (2021): "rwolf2 Implementation and Flexible Syntax," *Unpublished*.

Clarke, D., J. P. Romano, and M. Wolf (2020): "The Romano-Wolf multiple-hypothesis correction in Stata," *The Stata Journal*, 20, 812–843.

Currie, J. and D. Almond (2011): "Human capital development before age five," in *Handbook of Labor Economics*, vol. 4 of *Handbook of Labor Economics*, 1315–1486.

Desiere, S., W. Vellema, and M. D'Haese (2015): "A validity assessment of the Progress out of Poverty Index (PPI)," *Evaluation and Program Planning*, 49, 10–18.

Doyle, O., C. P. Harmon, J. J. Heckman, and R. E. Tremblay (2009): "Investing in early human development: Timing and economic efficiency," *Economics & Human Biology*, 7, 1–6.

Grantham-McGregor, S., A. Adya, O. Attanasio, B. Augsburg, J. Behrman, B. Caeyers, M. Day, P. Jervis, R. Kochar, P. Makkar, C. Meghir, A. Phimister, M. Rubio-Codina, and K. Vats (2020): "Group Sessions or Home Visits for Early Childhood Development in India: A Cluster RCT," *Pediatrics*, 146, e2020002725.

Innovations For Poverty Action (2016): "PPI for Tanzania 2011," *Report*.

Lee, D. S. (2009): "Training, Wages, and Sample Selection: Estimating Sharp Bounds on Treatment Effects," *Review of Economic Studies*, 76, 1071–1102.

Lu, C., M. M. Black, and L. M. Richter (2016): "Risk of poor development in young children in low-income and middle-income countries: an estimation and analysis at the global, regional, and country level," *The Lancet Global Health*, 4, e916–e922.

McCoy, D. C., G. Fink, and M. Waldman (2018a): "CREDI Data Management & Scoring Manual," *User guide*.

McCoy, D. C., C. R. Sudfeld, D. C. Bellinger, A. Muhihi, G. Ashery, T. E. Weary, W. Fawzi, and G. Fink (2017): "Development and validation of an early childhood development scale for use in low-resourced settings," *Population Health Metrics*, 15, 1 – 18.

McCoy, D. C., M. Waldman, CREDIFieldTeam1, and G. Fink (2018b): "Measuring early childhood development at a global scale: Evidence from the Caregiver-Reported Early Development Instruments," *Early Childhood Research Quarterly*, 45, 58 – 68.

Ozler, B., L. C. H. Fernald, P. Kariger, C. McConnell, M. Neuman, and E. Fraga (2018): "Combining pre-school teacher training with parenting education: A cluster-randomized controlled trial," *Journal of Development Economics*, 133, 448 – 467.

Richter, L. M., B. Daelmans, J. Lombardi, and e. al. (2017): "Investing in the foundation of sustainable development: pathways to scale up for early childhood development," *The Lancet*, 389, 103 – 118.

Romano, J. P. and M. Wolf (2005a): "Exact and Approximate Stepdown Methods for Multiple Hypothesis Testing," *Journal of the American Statistical Association*, 100, 94–108.

——— (2005b): "Stepwise Multiple Testing as Formalized Data Snooping," *Econometrica*, 73, 1237–1282.

Sabol, T. J., D. McCoy, K. Gonzalez, L. Miratrix, L. Hedges, J. K. Spybrook, and C. Weiland (2022): "Exploring treatment impact heterogeneity across sites: Challenges and opportunities for early childhood researchers," *Early Childhood Research Quarterly*, 58, 14–26.

Save the Children (2018): "Tuwekeze Pamoja Baseline Report," Tech. rep.

——— (2020): "Tuwekeze Pamoja: Snapshot of Midline Findings," Tech. rep.

Tauchmann, H. (2018): "Lee (2009) Treatment-Effect Bounds for Nonrandom Sample Selection," *The Stata Journal*, 14, 884–894.

UNICEF (2011): "Violence Against Children in Tanzania: Findings from a National Survey 2009," Tech. rep., Dar es Salaam, Tanzania.

A Data appendix

A.1 Constructed variables

A.1.1 Home environment

Table 13 lists the 9 categories of play material that are included in the home environment index. At baseline, children had access to on average 2.5 types of play materials; by the follow-up, this had increased 5.2 overall (4.9 in the control group); in part reflecting the greater age of the children.

Table 13: **Home environment: play material categories**

Homemade toys? (such as ball, stuffed dolls, cars, or other toys made at home out of local materials, slippers, clay etc.)
Toys from a shop or manufactured toys? (such as car, ball, animal, doll)
Household objects? (such as bowls, cups or pots)
Objects found outside? (such as sticks, stones or leaves)
Drawing or writing materials?
Any puzzles (even a two-piece puzzle counts)?
Anything that consists of two or three-pieces? (such as airplanes made of sticks and leaves or metal wheel with a stick)
Objects that teach about colors, sizes or shapes?
Objects or games that help teach about numbers/counting?

A.1.2 Caregiver practices

The respondent was asked how often, in the past week, either the mother, father or other caregiver did any of nine different activities with the child. Table 14 lists the full set of questions. The possible responses were: once; a few times (two or three); frequently (more than 3 times).

Table 14: **Caregiver-child interactions**

Read books or look at pictured books with the child?
Tell stories to the child?
Sing songs to or with the child, including lullabies/rhymes?
Take the child outside the home? For example, to the market, visit relatives.
Play any simple games with the child?
Name objects or draw things for or with the child?
Show or teach your child something new, like teach a new word, or teach them how to do something? (e.g. to hold a spoon)
Teach alphabet or encourage the child to learn letters?
Play a counting game or teach numbers to the child?

The respondent was asked how often, in the past week, either the mother, father or other caregiver tried one of the list of parenting strategies shown in Table 15. The possible responses at wave 2 were: never; once a month; sometimes (more than once to 5 times per

month); many times per month (more than 5 time but less than once a day); at least once a day. A similar, but slightly shorter, list of frequency responses was used at wave 1.

Table 15: **Positive and negative parenting**

Positive Parenting
Show affection to your child? (such as hug, hold closely, tickle their cheek, putting the child on your lap, kiss on the forehead and cheeks)
Tell the child that you love him/her?
Gave him/her something else to do?
When your child misbehaves do you explain why the behaviour was wrong?
Praised or encouraged your child?
Give the child a special privilege or reward?
Use rules to encourage your child to behave well?
Listen to what your child thinks?

Negative parenting
Speak negatively/unkindly to the child?
Yell / shout at your child for misbehaving?
Shake him/her?
Spank, hit or slap your child for misbehaving?
Hit multiple times, on the bottom or elsewhere on the body with something like a belt, stick or other hard object, your child for misbehaving?
Take away something they liked/wanted or forbid them to leave the house/do an activity?

A.2 Balance and attrition tables

A.2.1 Balance

Of a total first wave sample of 2,289 children, 1,153 (50.4%) were in the treatment group and 1,136 (49.6%) in control. Table 16 shows the balance of pre-treatment characteristics between these two groups. Treatment and control groups are well balanced in terms of demographic characteristics. There are statistically significant differences on gender of the primary caregiver (3% less likely to be female in control) and the probability that the sample child is the caregiver's first child (4% less likely in control); however, these differences are small in magnitude. Child sex and age, parental literacy, SES and mother's age are all well balanced.

The balance on pre-treatment values of the outcome variables is more nuanced. Child development scores are very well balanced, with a statistically insignificant difference of 0.02 standard deviations between treatment and control. Summary scores of home environment and caregiver practices show a consistent imbalance, however, with respondents in the treatment group reporting systematically higher values on all measures. These differences are statistically significant, and range from small to modest in magnitude (0.08%-0.24% of the control group standard deviation). It is not clear why this should be the case, but it suggests that there may be differences in response patterns between the two groups, particular for subjective responses such as these. It is worth noting that this applies both to measures of

good practices (e.g. 'positive parenting'), as well as to practices that the programme sought to discourage, such as those grouped into the 'negative parenting' summary variable, suggesting that social desirability bias on the part of caregivers who are aware of the upcoming intervention is not the full explanation.

Table 16: **Balance: treatment vs control at wave 1**

	(1) Control		(2) Treat		T-test Difference	Normalized difference
Variable	N	Mean/SE	N	Mean/SE	(1)-(2)	(1)-(2)
Girl child	1136	0.527	1153	0.509	0.018	0.036
		(0.015)		(0.015)		
Age in months	1136	7.602	1153	7.667	-0.065	-0.024
		(0.074)		(0.087)		
Female caregiver	1136	0.934	1153	0.963	-0.029***	-0.130
		(0.007)		(0.006)		
Mother literate	1136	0.832	1151	0.831	0.000	0.001
		(0.011)		(0.011)		
Father literate	1134	0.919	1137	0.908	0.011	0.040
		(0.008)		(0.009)		
Young mother	1136	0.527	1153	0.558	-0.030	-0.061
		(0.015)		(0.015)		
SES (in sd)	1136	-0.000	1153	-0.050	0.050	0.049
		(0.030)		(0.030)		
First child	1136	0.305	1152	0.344	-0.039**	-0.084
		(0.014)		(0.014)		
CREDI score	1136	0.032	1153	0.056	-0.024	-0.026
		(0.026)		(0.029)		
Kinds of toys	1135	0.000	1153	0.146	-0.146***	-0.139
		(0.030)		(0.032)		
Mother Interaction	1136	0.000	1153	0.237	-0.237***	-0.196
		(0.030)		(0.041)		
Father Interaction	1136	-0.000	1153	0.081	-0.081*	-0.073
		(0.030)		(0.036)		
Mother PP	1135	0.000	1151	0.199	-0.199***	-0.192
		(0.030)		(0.031)		
Mother NP	1135	0.000	1152	0.110	-0.110**	-0.105
		(0.030)		(0.032)		
Father PP	1129	-0.000	1135	0.121	-0.121***	-0.120
		(0.030)		(0.030)		
Father NP	1128	0.000	1131	0.238	-0.238***	-0.203
		(0.030)		(0.039)		

Notes: The value displayed for t-tests are the differences in the means across the groups. ***, **, and * indicate significance at the 1, 5, and 10 percent critical level.

A.2.2 Attrition

Table 17 compares attriters and panel members. The attrited observations are somewhat different from panel members: they are more likely to be first children and to be children

of young mothers, and less likely to be girls or children of literate mothers. The first two differences are the largest in magnitude, with attrited 25% (12 pp) more likely to be first born, and 20% (10 pp) more likely to be children of young mothers.

While attrition levels are quite similar across treatment and control groups, Table 18 shows that there are some differences in characteristics of those who drop out of the sample from treatment and control, respectively. Attriters in the treatment group are slightly older, more likely have a young mother and be first born. In contrast they are less likely to have literate parents, and come from lower SES households. This suggest that attrition in the control group is more positively selected, with older, more established, more educated and wealthier households dropping out, as compared with attriters in the treatment group.

While child development scores are well-balanced across attrited observations in treatment and control, there are also some marked differences in the caregiver practice variables. With the exception of the home environment variable (which shows no difference), the home and caregiver variables mirror the trend present across treatment and control groups at baseline, with attriters from the treatment group having higher values across the board - although the difference is not always statistically significant.

A.3 Further descriptive statistics

Table 19 gives the pairwise correlations between the seven constructed variables of home environment and caregiver practices. While some of the high correlations are not surprising (e.g. between mother and father measures within the household - recall these are reported by the same person), others are noteworthy. The variety of toys in the house has quite a low correlation with the behavioural measures, suggesting it is not simply proxying for the same set of engaged caregiver characteristics. Most striking is the fact that positive and negative parenting practices are positively correlated, although weakly so. This may be capturing that a general "propensity to respond positively," as suggested by a comparison of treatment and control groups at baseline on these variables. It may also suggest an underlying "propensity to engage" with the child, either in a disciplinary or affectionate way: those parents who are very engaged may practice more of both.

Table 20 compares respondents in the treatment group who did or did not attend at least one caregiver training session. The table suggests that there are not substantial differences between the two groups.

Table 17: **Attrition: attrited vs panel at wave 1**

Variable	(1) Panel N	Mean/SE	(2) Attrit N	Mean/SE	T-test Difference (1)-(2)	Normalized difference (1)-(2)
Girl child	1721	0.534	568	0.470	0.064***	0.128
		(0.012)		(0.021)		
Age in months	1721	7.602	568	7.737	-0.135	-0.050
		(0.066)		(0.114)		
Female caregiver	1721	0.949	568	0.947	0.002	0.008
		(0.005)		(0.009)		
Mother literate	1719	0.840	568	0.806	0.034*	0.090
		(0.009)		(0.017)		
Father literate	1710	0.915	561	0.907	0.008	0.028
		(0.007)		(0.012)		
Young mother	1721	0.518	568	0.616	-0.098***	-0.196
		(0.012)		(0.020)		
SES (in sd)	1721	-0.031	568	-0.006	-0.025	-0.025
		(0.024)		(0.045)		
First child	1721	0.295	567	0.414	-0.120***	-0.256
		(0.011)		(0.021)		
CREDI score	1721	0.049	568	0.031	0.018	0.019
		(0.022)		(0.040)		
Kinds of toys	1720	0.065	568	0.100	-0.034	-0.033
		(0.024)		(0.049)		
Mother Interaction	1721	0.108	568	0.156	-0.048	-0.040
		(0.029)		(0.053)		
Father Interaction	1721	0.019	568	0.108	-0.089	-0.080
		(0.025)		(0.056)		
Mother PP	1719	0.093	567	0.122	-0.029	-0.028
		(0.025)		(0.042)		
Mother NP	1720	0.072	567	0.005	0.067	0.064
		(0.025)		(0.044)		
Father PP	1702	0.074	562	0.019	0.056	0.056
		(0.024)		(0.044)		
Father NP	1698	0.122	561	0.111	0.011	0.009
		(0.028)		(0.051)		

Notes: The value displayed for t-tests are the differences in the means across the groups. ***, **, and * indicate significance at the 1, 5, and 10 percent critical level.

Table 18: **Attrited observations: treatment vs control at wave 1**

Variable	N	(1) Control Mean/SE	N	(2) Treat Mean/SE	T-test Difference (1)-(2)	Normalized difference (1)-(2)
Girl child	286	0.503	293	0.437	0.067	0.133
		(0.030)		(0.029)		
Age in months	286	7.768	293	8.463	-0.695**	-0.179
		(0.153)		(0.280)		
Female caregiver	286	0.937	293	0.959	-0.022	-0.099
		(0.014)		(0.012)		
Mother literate	286	0.839	293	0.768	0.071**	0.179
		(0.022)		(0.025)		
Father literate	285	0.930	287	0.885	0.045*	0.154
		(0.015)		(0.019)		
Young mother	286	0.577	293	0.645	-0.068*	-0.140
		(0.029)		(0.028)		
SES (in sd)	286	0.089	293	-0.103	0.191**	0.178
		(0.064)		(0.062)		
First child	286	0.367	292	0.449	-0.081**	-0.166
		(0.029)		(0.029)		
CREDI score	286	0.058	293	0.017	0.041	0.043
		(0.052)		(0.061)		
Kinds of toys	286	0.106	293	0.096	0.009	0.008
		(0.065)		(0.070)		
Mother Interaction	286	0.048	293	0.261	-0.214**	-0.169
		(0.062)		(0.084)		
Father Interaction	286	0.094	293	0.142	-0.048	-0.035
		(0.072)		(0.086)		
Mother PP	286	0.037	292	0.230	-0.193**	-0.191
		(0.060)		(0.059)		
Mother NP	286	-0.045	292	0.048	-0.092	-0.089
		(0.061)		(0.060)		
Father PP	284	-0.024	289	0.081	-0.106	-0.102
		(0.062)		(0.061)		
Father NP	283	0.009	289	0.207	-0.198**	-0.165
		(0.066)		(0.075)		

Notes: The value displayed for t-tests are the differences in the means across the groups. ***, **, and * indicate significance at the 1, 5, and 10 percent critical level.

Table 19: **Correlation of home and caregiver variables at wave 1**

Variables	Kinds of toys	Mother Interaction	Father Interaction	Mother PP	Mother NP	Father PP	Father NP
Kinds of toys	1.0000						
Mother Interaction	0.2727	1.0000					
	(0.0000)						
Father Interaction	0.2596	0.4992	1.0000				
	(0.0000)	(0.0000)					
Mother PP	0.1536	0.2821	0.1865	1.0000			
	(0.0000)	(0.0000)	(0.0000)				
Mother NP	0.1293	0.1207	0.0717	0.1955	1.0000		
	(0.0000)	(0.0000)	(0.0007)	(0.0000)			
Father PP	0.1472	0.2236	0.3027	0.7475	0.1567	1.0000	
	(0.0000)	(0.0000)	(0.0000)	(0.0000)	(0.0000)		
Father NP	0.1301	0.0872	0.1289	0.1307	0.6484	0.1997	1.0000
	(0.0000)	(0.0000)	(0.0000)	(0.0000)	(0.0000)	(0.0000)	

Notes: obs=2,254.

Table 20: **Wave 1 characteristics by attendance (treatment group only)**

	(1) Did not attend		(2) Attended		T-test Difference	Normalized difference
Variable	N	Mean/SE	N	Mean/SE	(1)-(2)	(1)-(2)
Girl child	439	0.526	431	0.541	-0.014	-0.029
		(0.024)		(0.024)		
Age in months	439	7.477	431	7.796	-0.320	-0.110
		(0.136)		(0.144)		
Female caregiver	439	0.959	431	0.970	-0.011	-0.058
		(0.009)		(0.008)		
Mother literate	437	0.844	431	0.856	-0.012	-0.033
		(0.017)		(0.017)		
Father literate	432	0.910	428	0.921	-0.011	-0.039
		(0.014)		(0.013)		
Young mother	439	0.510	431	0.541	-0.030	-0.061
		(0.024)		(0.024)		
SES (in sd)	439	0.016	431	-0.084	0.100	0.099
		(0.048)		(0.049)		
First child	439	0.330	431	0.281	0.050	0.107
		(0.022)		(0.022)		
Overall	439	0.111	431	0.033	0.078	0.080
		(0.051)		(0.041)		
Kinds of toys	439	0.124	431	0.203	-0.080	-0.076
		(0.050)		(0.051)		
Mother Interaction	439	0.247	431	0.213	0.033	0.024
		(0.060)		(0.071)		
Father Interaction	439	0.056	431	0.075	-0.019	-0.017
		(0.049)		(0.058)		
Mother PP	438	0.252	431	0.140	0.112	0.103
		(0.053)		(0.051)		
Mother NP	439	0.164	431	0.092	0.072	0.064
		(0.055)		(0.052)		
Father PP	428	0.168	428	0.109	0.059	0.059
		(0.050)		(0.046)		
Father NP	427	0.258	425	0.235	0.023	0.017
		(0.067)		(0.061)		

Notes: The value displayed for t-tests are the differences in the means across the groups. ***, **, and * indicate significance at the 1, 5, and 10 percent critical level.

B Cost effectiveness calculations

Calculating the cost effectiveness of a multifaceted intervention is challenging, even more so in a pilot study such as this, where a considerable share of the costs arise from the development of the intervention itself, monitoring & evaluation and capacity development. When fully implemented, Tuwekeze Pamoja works with caregivers, communities, schools and officials to ensure children are supported from conception through to the first years of primary school. The present paper is focused on evaluating a subset of the full intervention (two years' implementation of those elements targeting caregivers) on a subset of beneficiaries (children aged 4-12 months at baseline). Three important areas of change are not captured in these estimates: changes at the policy level in Tanzania (either locally or nationally) that affect early child development programming beyond the intervention; any effects of the intervention on caregivers' subsequent children (or indeed on caregivers themselves); and medium to longer term effects of the programme on the target children. With these limitations in mind, it is still important to document the costs of the project, and estimate the marginal cost, for comparability with other interventions.

The full cost of Tuwekeze Pamoja over five years (with approximately four years of active intervention, in addition to programme development and training) was budgeted for £2,5 million. The first three years of the project, which are those covered by this study, cost just under £1,4 million.[10] Of this, £259,499 are direct project costs (e.g. material development, consumables, community facilitator stipend, advocacy, technical support - but *excluding* salaries, monitoring & evaluation, overheads, capacity development and capital costs); £97,948 (38%) of which are attributed specifically to the Core caregiver-focussed activities, with a further £15,830 (6%) attributed to additional activities in the Core Plus treatment arm.

Over the full five years, the intervention is expected to directly reach 13,960 children, 12,642 caregivers, and 88 pre-primary and head teachers, as well as carry out advocacy at local, regional and national levels. The first three years of the project included two years of implementation. Over this time two cycles of caregiver training were run in each treatment village, with each cycle including one round of 0-3 sessions, and one round of 4-6 sessions. These sessions registered 6,850 caregivers, who collectively had 7,102 children

[10] Budget reporting for the project is in GBP. For the purposes of these calculation, actual spend is converted to 2017 GBP, using January values of the Retail Price Index of the Office of National Statistics (https://www.ons.gov.uk/economy/inflationandpriceindices).

Table 21: **Calculating costs per child**

Costs per child		Y1-Y3 reach		
		Total child reach	Children attended	2 x children attended
Costs		13,960	6,118	12,236
AC1: 5 year budget	£2,500,000	£179.08		
AC2: 3 year actual	£1,397,190		£228.37	£114.19
MC1: 3 year direct	£259,499		£42.42	£21.21
MC2: 3 year Core & Core+	£113,778		£18.60	£9.30

Notes: AC is average cost and MC is marginal cost. Four different costing approaches are described in the text: 5 year budget and total child reach are estimates from planning stage; all other figures are actual. Per child costs are total costs divided by reach. Cost and reach data are from the project team.

aged 0-6; of these, 5,870 caregivers and 6,118 children attended at least one training session. How many children age 0-6 are in the treatment area overall? This data is not available; however, we know that half of caregivers in the treatment sample attended at least one session; a reasonable estimate of the total number of children aged 0-6 in the treatment area is therefore double the number who attended at least one session (2 x 6,118 = 12,236).

To estimate cost effectiveness, we first calculate the per-child cost of the intervention, both on average and at the margin. This is done using three estimates of project reach (the number of beneficiaries): the expected reach, the actual reach (number who attended sessions), and the estimated total number of eligible children in the area. Given that the primary specification of this paper adopts an intention to treat approach, this suggests using the total number of children in the study area as the number of treated. For comparison, cost effectiveness is also calculated using the average treatment effect on the treated estimated, in which case the reach is the actual number of children who participated. An estimate of average cost based on total project budget and expected reach is also included.[11]

Table 21 summarises the costs, both overall and per child. Two approaches to average cost are presented: the first (AC1) divides the full programme budget by the estimated number of children who will be reached over the course of the project. The second (AC2) is more narrowly focussed on the first three years: two average costs are calculated from this, one per child who attended and one per child in the study area. Similarly, two sets of marginal costs are calculated: one assuming that all direct programme costs are the relevant measure

[11] Note that the estimated treatment effects are derived from the 0-3 caregiver programme exclusively, as the panel children were under three years old at both survey waves. Approximately half the children were in the 0-3 range, and half the caregiver training sessions were for this range. An alternative approach to costing the intervention elements evaluated in this paper would be to consider only half the costs of the implementation (an approximation of the 0-3 share) and only half the reach (an approximation of the number of children 0-3). This would lead to identical costs per child and cost effectiveness estimates.

Table 22: **Cost per standard deviation improvement: average and marginal costs**

		ITT	ToT
Cost approach		0.26	0.51
AC1	£179.08	£688.78	£351.14
AC2: ITT	£114.19	£439.18	
AC2: ToT	£228.37		£447.79
MC1: ITT	£21.21	£81.57	
MC1: ToT	£42.42		£83.17
MC2: ITT	£9.30	£35.76	
MC2: ToT	£18.60		£36.47

Notes: AC is average cost and MC is marginal cost. Four different costing approaches are described in the text and calculated in Table 21. Intention to Treat (ITT) cost effectiveness estimates use the primary ITT treatment effect estimate, and the estimated number of children in the treatment area as the reach estimate. Average Treatement effect of the Treated (ToT) cost effectiveness estimates use the ToT (IV) treatment effect estimate, and the number of children who attended sessions as the reach estimate.

of variable costs (MC1), the second counting only those programme costs specifically linked to the Core and Core + programming (MC2). The later excludes any direct programme costs associated with advocacy and the schools component of the programme, as well as technical assistance, travel and subsistence which apply to all programme elements.

To estimate cost effectiveness, we divide the relevant cost by the estimated treatment effect, to obtain a cost per standard deviation improvement in child outcomes. These estimates of cost effectiveness are shown in Table 22. Besides the first row, calculations are made separately for intention to treat (ITT) estimates and treatment effect on the treated estimates (ToT). The ITT cost effectiveness estimates relies on per child costs using the estimated number of children in the treatment area, combined with the ITT treatment effect as the denominator; the ToT estimates use the number of children who attended, combined with the ToT (IV) treatment effect.

It is clear from Table 22 that, if the only outcome of the intervention is the increase in early child development scores measured in this study, the average cost of the programme per unit of effect is very high. Even assuming that each child reached by the intervention saw a development gain equal to the ToT estimate, the average cost per standard deviation increase in development scores is over £350. Given the points made above, this estimate is likely extremely conservative; however, it does highlight that the logistics of setting up and running a complex intervention of this sort are non trivial.

This is further brought to light by the much lower estimates of marginal cost effectiveness. By the more conservative estimate, taking all direct programme costs into account, the marginal cost effectiveness is just over £80 per standard deviation. This figure includes a number of costs which would not vary with a very small expansion of the programme - such as technical assistance. The second marginal cost, using only those direct costs attributed to the caregiver training aspect of the intervention, is likely the best estimate of the cost of treating *one additional child* via the caregiver training programme. Based on this, the marginal cost effectiveness is £36 per standard deviation increase in early child development.[12]

[12]The cost effectiveness estimates other than AC1 (which uses the same reach figure for both ITT and ToT) are very similar. This is expected, as the reach figure for ITT is exactly twice that for ToT, by assumption, and the estimated ToT treatment effect is approximately twice as large as the ITT.