World Development 170 (2023) 106335

Contents lists available at ScienceDirect

World Development

journal homepage: www.elsevier.com/locate/worlddev

Fostering early childhood development in low-resource communities: Evidence from a group-based parenting intervention in Tanzania ☆

Margaret Leighton [a,*], Anitha Martine [b], Julius Massaga [b]

[a] University of St Andrews Castlecliffe, The Scores, St Andrews KY16 9AR, UK
[b] Save the Children International Tanzania, Tanzania

ARTICLE INFO

Article history:
Accepted 21 June 2023

JEL codes:
J13
I21
I25
I28
J18

Keywords:
Early child development
Child care policy
Parenting
Impact evaluation
Tanzania

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 a practice-led intervention piloted in rural Tanzania evaluated through a matched control study design. The core of the programme is an 8–10 week caregiver training course led by local facilitators, built around early stimulation and nurturing care. After two years of implementation, the intervention led to improvements in the development of 3-year olds of 0.29 standard deviations. Detailed data on caregivers indicates that these improvements are due to changes in the type and frequency of caregiver-child interactions for both mothers and fathers, as well as the quality of play materials in the home.

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, Harmon, Heckman, & Tremblay, 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 (LMICs) are at risk of failing to meet their developmental potential, with rates as high as 66% in Sub-Saharan Africa (Lu, Black, & Richter, 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 developmental delays from taking root (Richter, Daelmans, & Lombardi, 2017).

Interventions targeting early child development are not new: an established evidence base, largely from high-income countries, has demonstrated the effectiveness of three particular strategies: intensive home visiting, high quality centre-based early childhood education, and nutritional supplements for pregnant women (Currie & Almond, 2011). These approaches are either resource intensive, or limited in their scope for addressing developmental

☆ 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, Damian Clarke, Deborah Cobb-Clark, Thomas Dohmen, David Escamilla-Guerrero, Jakob Hennig and David Jaeger, as well as seminar participants at the University of Essex Institute for Social and Economic Research, the University of Glasgow and the University of St Andrews for helpful comments and discussions. Comments from the editor and two anonymous referees substantially improved the paper. Research assistance by Lauren Davis, Lawrence Ho, Himangshu Kumar and Peter Lowley 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.

* Corresponding author.
E-mail address: mal22@st-andrews.ac.uk (M. Leighton).

https://doi.org/10.1016/j.worlddev.2023.106335
0305-750X/

delays. In LMICs, where the need for such programmes is large but resources are limited, group-based parenting interventions are gaining in popularity. Studies from India (Grantham-McGregor et al., 2020) and Kenya (Luoto et al., 2021) suggest that group-based parenting interventions can achieve similar results as alternative models featuring one-to-one contact; contrasting evidence from China, however, finds that home visiting substantially outperforms group sessions (Sylvia et al., 2022). By training the adults who raise children, such programmes have the potential to (indirectly) reach the youngest children; the group model can achieve this at a relatively low cost. An effective, scaleable model of group-based caregiver training, which can be adapted to different contexts, would be valuable asset in the global effort to achieve SDG 4.2.

In 2017, Save the Children initiated a pilot study of a caregiver-focused 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-focused 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 in terms of the primary outcome variable and on objective demographic criteria, but not on all characteristics of interest: in particular, differences were revealed in baseline levels of home environment and caregiver practices, and on measures of remoteness.

A comparison of these two groups, controlling for baseline characteristics, shows that the intervention had a substantial positive effect on early child development. Intention to treat estimates show an improvement in child development of 0.29 standard deviations, while instrumental variables estimates of the effect of the treatment on the treated are considerably higher at 0.55 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 £32-£77 per standard deviation increase in child development.

While the core intervention shows substantial improvements in child development, there is no difference in the primary treatment effect across the two treatment arms. An exploration of the effect of the intervention on violent behaviours towards children and spouses also does not find any difference across the two treatments. It is important to note the relatively short timeline when interpreting these results; it could also be that changes of this type require a more intense intervention style. Whatever the cause, after two years, the additional violence-reduction intervention elements have not shown 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 (+0.25 sd), and both mothers and fathers had more frequent and varied interactions with the child (+0.37 sd and + 0.34 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).

Our study makes three contributions to the literature on early child development interventions. First, we provide new evidence on the effectiveness of a practice-led caregiver-training programme in a particularly poor and remote area. This intervention is a local adaptation of a set of approaches with a history of implementation by a major international organisation. As such, a particular contribution of our study is context-specific evidence on an implementation designed and carried out by the third sector. In contrast to research-led studies, which typically benefit from design input by world-leading experts, as well as a degree of monitoring which is difficult to maintain at scale, this evaluation comes closer to replicating how such interventions might look if rolled out on a regional or national level. With this in mind, it is striking to note that the impact of this intervention on domain-level child development measures, as well as caregiver behaviour, fall remarkably close to the average findings from a recent systematic review of rigorously-evaluated parenting interventions (Jeong et al., 2021).[1] This is an encouraging finding, as it suggests this model of intervention has the potential to scale effectively.

Second, this study helps fill the gap in evidence around such programmes in Sub-Saharan Africa. While parent training programmes have gained considerable popularity across Asia and Latin America, relatively few have been evaluated in Sub-Saharan Africa. In their review, Jeong et al., 2021 find only 9 randomised controlled trials with at least one site in an African country (of which just 6 included an element of group-based delivery). The same review identified 8 trials in Bangladesh alone, and 6 in the small island country of Jamaica, out of a total of 102 interventions. Recent evidence suggests that group caregiver training has considerable potential in Sub-Saharan Africa: in rural Kenya, Luoto et al. (2021) find that group-only training is just as effective, and considerably more cost-effective, than a model including home visits. Our study, with primarily group-based delivery, re-enforces these results: while the intention-to-treat effect sizes we estimate are a little smaller than Luoto et al. (2021), our findings from a different country context suggest that the effectiveness of such programmes can generalise.

Finally, the focus on fathers in our study, both in the design of the intervention and in the data collection and analysis, sets it apart from the majority of studies in this area. In a 2014 global evidence review, Panter-Brick et al. (2014) highlight the negligible presence of fathers in the parenting intervention literature. In their more recent review of parent training interventions, Jeong et al., 2021 find that only 7% of interventions engaged fathers in any way, with just one study measuring outcomes from fathers directly. Including fathers in early child development programmes has implications for children themselves, and also for gender

[1] Jeong et al., 2021 report standardized mean differences of 0.24 for motor, 0.32 for cognitive, 0.28 for language and 0.19 for socio-emotional development. These compare to our point estimates of 0.26 for motor, 0.31 for cognitive, 0.27 for language and 0.28 for socio-emotional (see Table 4). Regarding caregivers, Jeong et al., 2021 estimate a mean effect of 0.39 on parent–child interactions and 0.33 on parenting practices. Our equivalents are 0.37 (mothers) and 0.34 (fathers) for parent–child interactions; for parenting practices our estimates range from 0.18 (fathers' positive parenting practices) to 0.41 (decrease in mothers' punishment practices).

stereotypes and division of labour within the home. While our study, similar to the one identified by Jeong et al., 2021, only directly surveyed a small fraction of fathers (those who identified as the child's primary caregiver), each wave of data collection elicited detailed information on the quantity, quality, and type of interaction the child has with both the mother and the father. This allows us to estimate the impact of the intervention on the behaviours of both parents. While the caregiver training sessions in our study were primarily attended by mothers, we nevertheless find substantial changes in fathers' interactions with children. This suggest important spillovers within the household: to the best of our knowledge, we are the first paper to show this quantitatively.

The remainder of the paper proceeds as follows. Section 2 describes the intervention and the study design. Section 3 describes the data collection, sample, variables of interest, 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. The present 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.[2]

The caregiver sessions were delivered by volunteer community facilitators. Prospective facilitators were identified by community leaders and government officials based on their skills (literacy, community mobilisation and facilitation experience) and interest; final selection was made after a short interview. Facilitators received a monthly stipend of 30,000 shillings (approximately £10.40 in 2017), and underwent an initial five day training. A refresher training (approximately one day of training for each 3–5 sessions) was also delivered at the start of each cycle.

Anecdotal evidence from the field suggests that the community facilitator model was particularly effective in this setting. The facilitators knew and understood the community, spoke the local language and appreciated the challenges caregivers faced. Their advice was therefore both more tailored to the local context, and also more readily accepted. Programme staff stressed the importance of status and recognition for motivating community volunteers, and highlighted the need for ongoing capacity building over the course of the intervention.

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 focused on violence-reduction (including conflict resolution, stress management, and gender and 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.

A particular feature of the intervention is the contextual adaptation which was embedded into its design. While the core content was drawn from existing, evidence-based approaches, the inception stages of the intervention adapted this material and drew up the curriculum. Two structural features of the adaptation and implementation should be noted. First, at the outset a detailed assessment of the local environment was carried out by a consultant: the information gathered in this way was then presented back to local stakeholders for their feedback and discussion. Both the analysis by the consultant and the feedback from the validation meeting were used to ensure the programme was targeted and relevant. Second, the programme was implemented by Save the Children in partnership with a local organisation. ADP Mbozi, the local partner, brought extensive experience working in nearby communities, as well as in programme design and delivery experience. This led to important adaptations of the model which might not arise in a one-size-fits-all-contexts approach, e.g. caregiver training sessions were scheduled during the dry season, as agricultural activities are at their peak when the rains come and availability of caregivers would be limited.

Once the intervention was underway, the programme team sought to remain responsive and adaptive. A review of the curriculum was scheduled after the first year of implementation. This review engaged caregivers who had attended sessions, as well as community facilitators who had delivered them. The findings allowed the team to refine the curriculum content and delivery, refocussing sessions on the most relevant content and, in some cases, reducing the time or frequency of sessions.

[2] 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.

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 by the implementation team 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:[3] 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. 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. Fig. 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 intervention are a panel of child-caregiver dyads, to be surveyed every two years over the life of the project. The panel is focused on a group of children who were infants when the project began. Since the intervention was focused 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.

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; 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 resurveyed 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), 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.

CREDI is a low-cost tool which has been shown to have good properties of validity and reliability (McCoy, Waldman, CREDIFieldTeam1, & Fink, 2018; Munoz-Chereau, Ang, Dockrell, Outhwaite, & Heffernan, 2021; McCoy, Seiden, Waldman, & Fink, 2021). In a systematic review of early child development measurement in LMICs, Munoz-Chereau et al. (2021) identify only four out of 31 total measures which are caregiver-reported and free to use. These criteria are particularly important in our context where data collection was to be carried out by non-specialist enumerators (ruling out direct assessments requiring particular expertise), and where the evaluation budget, as a share of the implementation budget, faced tight constraints. Finally, among these four tools, CREDI demonstrates the widest range of translations and local adaptations (Munoz-Chereau et al., 2021, Table 2). In particular, one of the first applications of CREDI was in Tanzania, for which the tool underwent careful translation by the original development team (McCoy et al., 2017).

As a relatively recent measure, the evidence base around CREDI is still developing, including how it compares with more established tools. While Alderman, Friedman, Ganga, Kak, and Rubio-Codina (2021) find that CREDI shows adequate validity compared to the (direct expert assessment) Bayley-III in India, Li, Tang, Bai, Zhao, and Shi (2020) find more mixed results in China. While the full sample correlation between CREDI and Bayley-III in Li et al. (2020)'s study was quite strong, it was much weaker in individual age brackets, with an initial moderate correlation (ages 5–18 months) becoming weak and inconsistent thereafter. This suggests that the validity of CREDI may decline with age, at least in some contexts. Given that our purpose in this study is to comparison two populations, rather than assess the overall level of development per se, we keep this limitation in mind but proceed cautiously with CREDI.

In practice, 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 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 set 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 (Schreiner, 2016); for a validity assessment of this measure see, e.g. Desiere, Vellema, & D'Haese (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. Characteristics of the sample

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.

[3] 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.

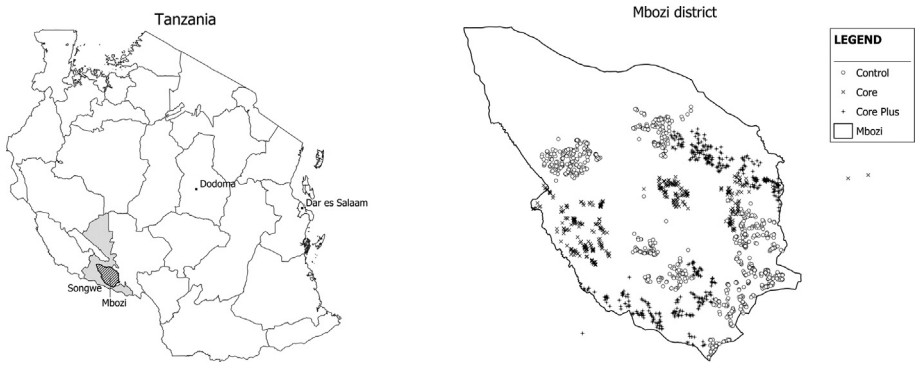

Fig. 1. Study area with Treatment and Control locations. (Maps plotting the study area in the context of Tanzania (L), and Treatment and Control locations within Mbozi district (R). Although Treatment and Control areas were assigned based on ward boundaries at the project inception in 2017, ward boundaries 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 fromhttps://data..humdata.org).

Table 1
Balance: treatment and control at wave 1 and panel.

		Wave 1					Panel			
		Control		Treat	p-value		Control		Treat	p-value
Variable	N	Mean/SE	N	Mean/SE	(C)-(T)	N	Mean/SE	N	Mean/SE	(C)-(T)
Girl child	1136	0.527	1153	0.509	0.384	849	0.535	873	0.545	0.662
		(0.015)		(0.015)			(0.017)		(0.017)	
Age in months	1136	7.602	1153	7.667	0.568	849	29.127	873	29.333	0.128
		(0.074)		(0.087)			(0.089)		(0.101)	
Female caregiver	1136	0.934	1153	0.963	0.002***	849	0.934	873	0.960	0.016**
		(0.007)		(0.006)			(0.009)		(0.007)	
Mother literate	1136	0.832	1151	0.831	0.979	845	0.854	874	0.847	0.652
		(0.011)		(0.011)			(0.012)		(0.012)	
Father literate	1134	0.919	1137	0.908	0.342	827	0.929	843	0.930	0.914
		(0.008)		(0.009)			(0.009)		(0.009)	
Young mother	1136	0.527	1153	0.558	0.145	836	0.457	870	0.449	0.756
		(0.015)		(0.015)			(0.017)		(0.017)	
SES (in sd)	1136	−0.000	1153	−0.050	0.239	850	−0.000	878	0.000	0.998
		(0.030)		(0.030)			(0.034)		(0.032)	
First child	1136	0.305	1152	0.344	0.045**	850	0.255	876	0.284	0.176
		(0.014)		(0.014)			(0.015)		(0.015)	
Dist to highway	1134	18.393	1097	9.558	0.000***	850	19.016	833	9.624	0.000***
		(0.303)		(0.200)			(0.339)		(0.228)	
Dist to any road	1134	1.586	1097	1.526	0.461	850	1.713	833	1.486	0.015**
		(0.059)		(0.057)			(0.069)		(0.062)	
Dist to nearest town	1134	23.148	1097	19.896	0.000***	850	23.997	833	20.178	0.000***
		(0.312)		(0.178)			(0.338)		(0.198)	
Overall	1136	0.032	1153	0.056	0.537	848	−0.084	868	0.204	0.000***
		(0.026)		(0.029)			(0.031)		(0.035)	
Kinds of toys	1135	0.000	1153	0.146	0.001***	848	0.000	878	0.282	0.000***
		(0.030)		(0.032)			(0.034)		(0.035)	
Mother Interaction	1136	0.000	1153	0.237	0.000***	852	0.000	880	0.469	0.000***
		(0.030)		(0.041)			(0.034)		(0.041)	
Father Interaction	1136	−0.000	1153	0.081	0.080*	852	−0.000	880	0.302	0.000***
		(0.030)		(0.036)			(0.034)		(0.043)	
Mother PP	1135	0.000	1151	0.199	0.000***	831	0.000	852	0.372	0.000***
		(0.030)		(0.031)			(0.035)		(0.042)	
Mother NP	1135	0.000	1152	0.110	0.012**	830	0.000	850	−0.339	0.000***
		(0.030)		(0.032)			(0.035)		(0.036)	
Father PP	1129	−0.000	1135	0.121	0.004***	799	−0.000	796	0.175	0.001***
		(0.030)		(0.030)			(0.035)		(0.042)	
Father NP	1128	0.000	1131	0.238	0.000***	803	−0.000	790	−0.205	0.000***
		(0.030)		(0.039)			(0.035)		(0.034)	

Notes: table shows baseline data for wave 1 sample (left) and panel observations (right). Standard errors in parentheses. P-value regards a t-test of equal means between control and treat: * $p < 0.10$ ** $p < 0.05$, *** $p < 0.01$.

As the treatment and control areas were not selected at random, there is a risk of imbalance in characteristics. Table 1 presents two sets of checks for balance across treatment and control: first, the full sample at baseline, and second, the baseline levels of panel observations. In terms of pre-determined socio-demographic characteristics, including child sex, age, parental literacy, and socioeconomic status, the treatment and control groups are well-balanced, both at baseline and within the panel. Balance on two other sets of baseline characteristics is more problematic. Measures of remoteness (distance to the local highway, to any road, and to nearest

Table 2
Normalised CREDI scores in Wave 2.

Variable	Control		Treat	
	Mean	Std. Dev.	Mean	Std. Dev.
Overall	−0.084	0.915	0.204	1.03
Motor	−0.072	1.055	0.178	1.171
Cognitive	−0.091	1.117	0.208	1.232
Language	−0.123	0.84	0.172	0.934
Socio-emotional	0.239	1.164	0.497	1.295
Sample	848		868	

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

town) indicate that the control observations live further from major infrastructure (highway and towns). Home and caregiver characteristics are also unbalanced, with the treatment group more likely to respond affirmatively to all questions. While baseline child CREDI scores are strongly balanced in the first wave, the panel shows a notable gap favouring the treatment group.

These imbalances must be accounted for in estimating the impact of the intervention. Our primary specification in all regressions therefore controls for baseline levels of socio-demographic, distance, home environment and caregiver practice variables. To explore the potential impact of imbalance, the main results are re-estimated using varying sets of controls. Finally, we check for heterogenous impacts of the intervention by interacting treatment with baseline measures.

The imbalance in baseline characteristics raises questions about the selection of treatment and control wards. As noted in Section 2.2, this was carried out by the implementation team, following some basic criteria: the implementation team wanted a set of treatment wards that spanned the diversity of the local area, but also required each ward to have a health centre. Control wards were chosen to match the treatment wards on geographic and socio-economic characteristics; however, lacking adequate mapping data this was done in an ad hoc way based on local knowledge.

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 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 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.

3.3. Child outcomes

The CREDI tool generates age-referenced development scores, normalised against a reference population. In the first wave, when the panel children were 4–12 months old, both the treatment and control groups scored on average very close to reference population (treat: +0.066 SD, control: +0.024 SD.). By the second wave, when the children were aged 2–3, differences emerge: the control group had now fallen slightly behind the reference population, scoring 0.089 standard deviations below the reference. The treatment group, on the other hand, had pulled ahead by 0.203 standard deviations.

Table 2 and Fig. 2 give an overview of wave 2 CREDI scores, both overall and by domain. The control group gives an indication of child development levels in the absence of the intervention. While the overall score, as well as three of the domains, indicate that the control group children are falling behind the reference group, two things are worth noting. First, the scores in the socio-emotional domain defy this trend, registering a substantial + 0.239 SD above the reference group. Second of all, these scores are considerably higher than those measured on a group of similarly-aged children in Zanzibar, Tanzania in 2019. In a cross-section of 499 children aged 18–29, Russell et al. (2022) find much larger developmental delays on CREDI domains, ranging from −0.494 on language to −0.116 on motor. In Russell et al. (2022)'s sample, 9.6% of children had an overall CREDI score more than 2 SD below the reference group, a range they define as representing a significant developmental concern; 18.4% had an overall score between 1 SD and 2 SD below the reference group (classed as a developmental concern). The equivalent figures from our control sample are 2.8% (24 out of 848) in the significant developmental concern group, and 10.5% (89 out of 848) in the developmental concern group - very similar shares to what we would expect in a normal distribution centred on the reference population mean. This suggests that, while there are signs of average developmental deficits in our control group by wave 2, the child development levels in our study area are considerably better than in some other areas of Tanzania.

3.4. 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 (diversity of toys), and six on caregiver-child interactions and parenting practices (three indices, computed separately 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 A Table 14). 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.[4] Nine activities were listed, including reading books, singing songs, playing games, and taking the child out of the home (see Appendix A Table 15 for the full list); most of these also include frequency data (from never, to more than three times in the last week).

[4] 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.

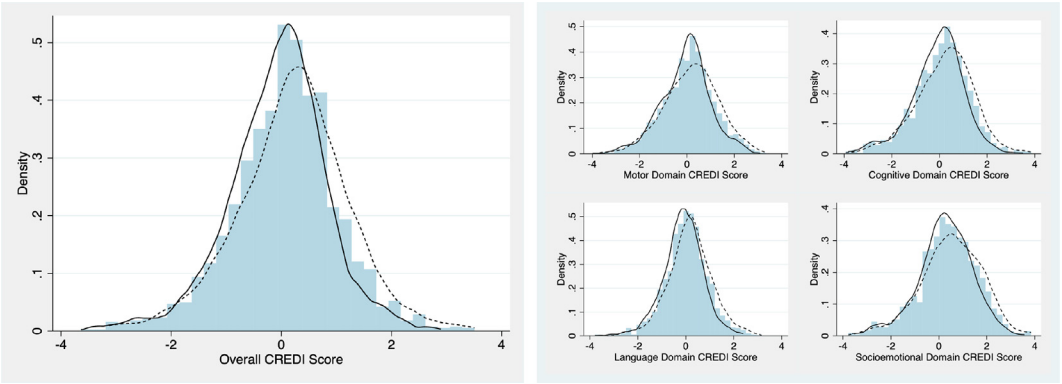

Fig. 2. Distribution of CREDI scores in Wave 2. (Notes: distribution of CREDI scores based on 1,716 wave 2 observations. Histogram bars show the full sample; solid line represents the control group and the dashed line represents treatment group. Graph display is truncated at ±4 SD for readability.).

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 focused 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 A Table 16.) 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.

3.5. 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; however, attrited observations had less remote locations at baseline.

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. There are differences in the remoteness of attrited individuals across treatment and control; most of those differences follow the average differences across the two groups, with the exception that attrited individuals from the control group lived closer to the nearest road.

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 three 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. Finally, the main results are re-estimated using propensity score matching, with varying degrees of trim.

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 suggests 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:

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

where Y_{it} is an individual outcome of interest measured at waves $t = 1, 2$; *treat* is a binary variable; and X_{i1} is a set of predetermined control variables. These controls, which are summarised in Table 1, include socio-demographic characteristics of the child and their parents, baseline values of primary and secondary outcome variables, and measures of remoteness. 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, Eq. 1 is modified to estimate the treatment effect in the two arms separately, as follows:

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

In Eq. 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. 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 (villages typically have 3–6 hamlets), 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.[5] If the effect of the intervention is heterogenous this would drive a further wedge between the ITT and ToT estimates. see Table 20.

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:

$$D_i = \beta_0 + \beta_1 \text{treat}_i + \gamma X_{i1} + \nu_{it} \quad (3)$$

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

where D_i is attendance at one or more caregiver training sessions and $treat_i$, treatment status, is the instrument excluded in Eq. 4. The other variables are as defined in Eq. 1. The coefficient of interest is $\hat{\alpha_{2SLS}}$, 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; 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 traditional thresholds of statistical significance 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.

We explore the robustness of our results to multiple-hypothesis corrections in two ways, applying the (highly conservative) Bonferroni correction and the (more powerful) Romano-Wolf correction. The Romano-Wolf correction, introduced by Romano and Wolf (2005) and Romano and Wolf (2005) 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. We apply the implementation of this method in Stata by Clarke, Romano, and Wolf (2020) and Clarke (2021). The corrected p-values are discussed along with the most relevant secondary results in Section 5 below; further details on our implementation of the corrections are presented in Appendix B.1. Note that we do not apply the correction to the extensions in Section 6, as we consider those results to be exploratory by design.

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 Eqs. 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 0.287, 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 28.7% 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 very similar point estimates (0.270 vs 0.306). This shows that the additional components of the Core Plus intervention did not substantially improve child development outcomes, with respect to the Core intervention package. Given this finding, the remainder of the analysis will focus primarily on a comparison of treatment (combining both arms) and control. We will consider the two separately when we look at the effect of the intervention on violence.

The models estimated in Table 3 also give 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 0.165 standard deviation higher CREDI scores

[5] 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.

Table 3
Child development: normed CREDI scores.

	(1) Overall		(2) Overall	
Treat	0.287***	(0.0772)		
Core			0.270***	(0.0685)
Core Plus			0.306***	(0.0901)
Overall baseline	0.165***	(0.0334)	0.164***	(0.0335)
First child	0.0836*	(0.0475)	0.0830	(0.0475)
Female caregiver	−0.0847	(0.110)	−0.0861	(0.109)
Girl child	0.0284	(0.0472)	0.0280	(0.0471)
Age in months	0.0148	(0.0113)	0.0148	(0.0113)
SES (in sd)	0.0610*	(0.0291)	0.0606*	(0.0293)
Young mother	−0.195***	(0.0618)	−0.193***	(0.0615)
Mother literate	0.158*	(0.0829)	0.156*	(0.0837)
Father literate	0.0866	(0.0834)	0.0872	(0.0835)
Dist to highway	0.00497	(0.00481)	0.00533	(0.00492)
Dist to any road	−0.0228	(0.0237)	−0.0239	(0.0238)
Dist to nearest town	−0.00433	(0.00567)	−0.00465	(0.00565)
Kinds of toys	0.0626**	(0.0290)	0.0633**	(0.0292)
Mother Interaction	0.0720***	(0.0206)	0.0716***	(0.0208)
Father Interaction	−0.0311	(0.0229)	−0.0314	(0.0230)
Mother PP	−0.0320	(0.0324)	−0.0322	(0.0326)
Mother NP	0.0248	(0.0292)	0.0242	(0.0291)
Father PP	0.0588*	(0.0278)	0.0587*	(0.0277)
Father NP	0.00414	(0.0226)	0.00424	(0.0224)
Constant	−0.226	(0.205)	−0.222	(0.205)
Observations	1638		1638	
F-T1vsT2			0.728	
F-pval			0.407	

Notes: dependent variable is in standard deviations of reference population. All control variables are shown. *PP* stands for positive parenting; *NP* stands for negative parenting. Standard errors in parentheses clustered at the ward level; * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

2 years later. This is similar in magnitude to the coefficient on mother's literacy (0.158), but considerably more than that on socio-economic status (SES): a 1 standard deviation increase in SES is associated with a 0.061 increase in CREDI scores. The indicator for having a young mother has the largest magnitude coefficient, associated with a −0.195 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, show no statistically significant association.

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. This allows us to explore the effect of the intervention on different aspect of early child development. Given that these subdomains were not the primary target of the intervention, we adjust our assessment of the statistical significance of these results to correct for multiple hypothesis testing. These corrections, which were applied simultaneously across the estimates presented in Table 4 (CREDI domains) and Table 5 (home environment and caregiver practices), are summarised in the tables, with full details in Appendix B.1.

Table 4 presents the estimated treatment effects of the intervention on the four domain scores. The estimated effects are all close to the overall estimate (which effectively is an average of the four): ranging from 0.26 (motor) to 0.31 (cognitive). It is interesting to note how similar our estimates are to the mean effects reported in Jeong et al., 2021's systematic review of randomised controlled trials of parenting interventions. Indeed, with the exception of socio-emotional (where we find a large effect of 0.28, compared with their estimate of 0.19), our estimated treatment effects are all within 0.02 SD of the effects they derive through meta-

Table 4
Child development subdomains: normed CREDI scores.

	(1) Motor	(2) Cognitive	(3) Language	(4) Socio-emotional
Treat	0.263	0.311	0.268	0.284
	(0.106)	(0.0909)	(0.0727)	(0.0900)
Controls	Yes	Yes	Yes	Yes
Unadjusted	**	***	***	***
Bonferroni		**	**	*
Romano-Wolf	*	**	**	*
Observations	1638	1638	1638	1638

Notes: dependent variable is in standard deviations of reference population. Control variables (not shown) include baseline value of the outcome variable, as well as all baseline child and caregiver controls (see Table 1). Standard errors in parentheses, clustered at the ward level. Statistical significance indicated based on individual regressions (Unadjusted), Bonferroni and Romano-Wolf (multiple hypothesis testing adjustments made across the 11 estimations shown in Tables 4 and 5). * $p < 0.10$ ** $p < 0.05$, *** $p < 0.01$.

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	0.253	0.365	0.339	0.306	−0.406	0.181	−0.233
	(0.0594)	(0.0599)	(0.0753)	(0.0954)	(0.0573)	(0.0580)	(0.0563)
Controls	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Unadjusted	***	***	***	***	***	***	***
Bonferroni	**	**	**	*	**	*	**
Romano-Wolf	**	**	**	*	***	*	**
Observations	1630	1636	1636	1592	1589	1518	1514

Notes: dependent variable is in standard deviations. Control variables (not shown) include baseline value of the outcome variable, as well as all baseline child and caregiver controls (see Table 1). *Int* stands for interactions; *PP* stands for positive parenting; *NP* stands for negative parenting. Standard errors in parentheses, clustered at the ward level. Statistical significance indicated based on individual regressions (Unadjusted), Bonferroni and Romano-Wolf (multiple hypothesis testing adjustments made across the 11 estimations shown in Tables 4 and 5); * $p < 0.10$ ** $p < 0.05$, *** $p < 0.01$.

analysis. The relatively large effect of the intervention on socio-emotional development is noteworthy: a recent review of early childhood development and parenting interventions in China finds that such programmes have smaller and less consistent impacts on socio-emotional development than on cognitive scores (Emmers et al., 2021).

These estimates are all highly statistically significant when evaluated individually; while they remain conventionally significant after the Romano-Wolf multiple hypothesis correction, for two of the domains the significance level drops to 10%. Overall, these results suggest that the intervention was effective at improving child development across this full range of domains. The relatively large coefficients on cognitive and socio-emotional development point to a particular impact in those areas.

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 Eq. 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. Corrections for multiple hypothesis testing have been applied simultaneously to these results and the CREDI domains presented in Table 4 (full details in Appendix B.1).

Table 5 shows the estimated treatment effects on these seven indicators. All show substantial changes in the expected direction: improvements in the diversity of play materials in the home, increases in the 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 0.41 to 0.18 standard deviations, with larger changes seen for mothers than for fathers. While the estimated treatment effects are all highly statistically significant when evaluated individually, the significance of the effects on mothers' and fathers' positive parenting practices drops to the 10% threshold after the Romano-Wolf correction.[6]

[6] Note that the sample size for some of these regressions is slightly smaller than for the child outcomes: while the interaction variables take a value of 0 if the parent never interacts with the child, the parenting practice variables would return a missing value. While most parents were living with the child in wave 1 (when they were young babies), more parents are absent in wave 2. In the data, an absent father would have 0 interactions with the child, but would have a missing value for positive and negative parenting practices. For that reason the sample sizes are smaller in columns (4)-(8), as the parenting practice data is missing for some number of parents.

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 experience 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 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.

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 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 Eqs. 3 and 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.

The high level of compliance with the intervention design is noteworthy. While we do not know which caregivers had been offered access to the intervention by the time of wave 2 data collection, we know roll-out was incomplete – and yet nearly 50% of respondents in treated areas reported some engagement with the sessions. This suggests there was strong motivation to participate among the study participants. It is also interesting that no members of the control group accessed the treatment. There were no rules preventing this; however, anecdotal evidence from the

Table 6
Predicting attendance.

	(1) Attended	
Treat	0.514***	(0.0767)
First child	−0.0430	(0.0290)
Female caregiver	0.0325	(0.0295)
Girl child	0.0130	(0.0213)
Age in months	0.00578	(0.00379)
SES (in sd)	−0.0161	(0.0109)
Young mother	0.0449	(0.0310)
Mother literate	0.0164	(0.0326)
Father literate	0.0203	(0.0390)
Dist to highway	−0.0000495	(0.00802)
Dist to any road	0.0000585	(0.0125)
Dist to nearest town	0.00331	(0.00729)
Kinds of toys	0.0167*	(0.00813)
Mother Interaction	−0.000386	(0.00829)
Father Interaction	−0.00555	(0.00999)
Mother PP	−0.0204	(0.0154)
Mother NP	−0.0170	(0.0109)
Father PP	0.0114	(0.0115)
Father NP	0.00168	(0.0114)
Constant	−0.203*	(0.105)
Observations	1652	
R^2	0.353	

Notes: regression of attendance (binary) on treatment status and other controls. All control variables are shown. *PP* stands for positive parenting; *NP* stands for negative parenting. Standard errors in parentheses clustered at the ward level; statistical significance is indicated based on model p-values; * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

field suggests that proximity to the session location was critical for participation. Given that no sessions were held in control group areas, it is perhaps not surprising that control group members did not attend.

Table 7 presents results from instrumental variables estimates of the primary and secondary 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 and 5: the estimated effect of the intervention on child development rises from 0.287 sd (OLS) to 0.553 sd (IV).

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

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	0.553***	0.492***	0.711***	0.660***	0.593**	−0.784***	0.347**	−0.447***
	(0.171)	(0.115)	(0.146)	(0.164)	(0.251)	(0.184)	(0.146)	(0.141)
Controls	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Observations	1638	1630	1636	1636	1592	1589	1518	1514

Notes: treatment effect estimates from 2SLS regressions where attendance is instrumented by treatment status. Control variables (not shown) include baseline value of the outcome variable, as well as all baseline child and caregiver controls (see Table 1). *Int* stands for interactions; *PP* stands for positive parenting; *NP* stands for negative parenting. Standard errors in parentheses clustered at the ward level; statistical significance is indicated based on model p-values; * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

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	0.303***	0.417	0.329***	0.330**	0.286***	0.281***	0.300*	0.120
	(0.0758)	(0.269)	(0.104)	(0.154)	(0.0767)	(0.0859)	(0.152)	(0.150)
Treat = 1 × First child	−0.0570							
	(0.102)							
Treat = 1 × Female caregiver		−0.136						
		(0.271)						
Treat = 1 × Girl child			−0.0770					
			(0.0877)					
Treat = 1 × Age in months				−0.00562				
				(0.0170)				
Treat = 1 × SES (in sd)					−0.0439			
					(0.0543)			
Treat = 1 × Young mother						0.0120		
						(0.121)		
Treat = 1 × Mother literate							−0.0156	
							(0.139)	
Treat = 1 × Father literate								0.184
								(0.141)
Controls	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Observations	1638	1638	1638	1638	1638	1638	1638	1638

Notes: outcome variable is in standard deviations of reference population. Control variables (not shown) include baseline value of the outcome variable, as well as all baseline child and caregiver controls (see Table 1). Standard errors in parentheses clustered at the ward level; statistical significance is indicated based on model p-values; * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

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 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 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 heterogeneity 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.

6.1.3. Violence in the home

Table 3 shows no difference in the effect of the two intervention arms on child development. Scoping work for the intervention had identified violence in the home, specifically the use of violent disciplinary measures towards wives and children, as a prevalent issue in the study area. Physical abuse, emotional abuse, and neglect in childhood are associated with a wide range of negative outcomes throughout the life cycle (Norman et al., 2012). Intimate partner violence (IPV) affects both the home environment and the welfare of primary caregivers. IPV is negatively associated with early child development across a range of LMICs (Jeong, Adhia, Bhatia, McCoy, & Yousafzai, 2020); recent research has confirmed this link in Tanzania as well (Oliveira et al., 2022). The Plus package of interventions was designed to address both IPV and violence (physical, emotional or neglect) towards children, in part because these may have a similar origin. In addition to five further caregiver training sessions, which covered topics including conflict resolution and stress management, the Plus package included community theatre events and engagement with local leaders.

The finding that this Plus package did not lead to any further improvements in child development beyond the Core intervention could be due to two things: either that the Plus package did not reduce violence, or that any reduction in violence that it did produce did not translate into improvements in child development. Given the complex relationship between violence and child development, there are many reasons for which the latter might hold, in particular over the relatively short two-year timeline of this study. To explore these two possibilities, we compare the effect of the Core and the Core Plus on two sets of outcomes: our index of positive and negative parenting practices, and responses to a question asked of caregivers in each wave: *In the past month, did you and your partner resolve a conflict violently?* The negative parenting practice index includes *'how often do you'* activities such as yell or shout at your child, hit your child, or spank your child for misbehaving (see Appendix A for the full list). The possible responses for the inter-partner question ranged from *never* to *at least once a day*: we consider two binary indicators, 'ever' and 'daily.'

Table 10 summarises these results. There is no difference in the treatment effect on positive and negative (disciplinary) parenting across Core and Core Plus. The point estimates for the reduction in negative parenting is substantially larger in Core than Core Plus for both mother and father, but the difference is not statistically significant. Compared to other outcomes, there is very little effect of the intervention on the share of caregivers who report using violence to resolve conflict with their partner. While the estimate shows a statistically significant reduction in the daily use of violence in Core areas, the effect size is small, and it is only weakly significant. In contrast, the point estimates for the Core Plus area are positive, although also small and statistically insignificant. The equality of the two treatment effects is only rejected in Column (5), with Core showing a statistically larger reduction in the use of violence ever; however the difference remains small and the finding should be taken as speculative.

Ultimately, we find no evidence that the Plus package reduced violence in the home any more than the Core intervention. The

Table 9
Child development: heterogeneity by targeted practices.

	(1) CREDI	(2) CREDI	(3) CREDI	(4) CREDI	(5) CREDI	(6) CREDI	(7) CREDI
Treat = 1	0.285***	0.285***	0.287***	0.289***	0.286***	0.287***	0.284***
	(0.0771)	(0.0783)	(0.0770)	(0.0779)	(0.0771)	(0.0773)	(0.0755)
Treat = 1 × Kinds of toys	0.0359						
	(0.0505)						
Treat = 1 × Mother Int		0.0378					
		(0.0263)					
Treat = 1 × Father Int			−0.0154				
			(0.0285)				
Treat = 1 × Mother PP				−0.0271			
				(0.0336)			
Treat = 1 × Mother NP					0.0572		
					(0.0411)		
Treat = 1 × Father PP						−0.00346	
						(0.0295)	
Treat = 1 × Father NP							0.0366
							(0.0455)
Controls	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Observations	1638	1638	1638	1638	1638	1638	1638

Notes: outcome variable is in standard deviations of reference population. Control variables (not shown) include baseline value of the outcome variable, as well as all baseline child and caregiver controls (see Table 1). *Int* stands for interactions; *PP* stands for positive parenting; *NP* stands for negative parenting. Standard errors in parentheses clustered at the ward level; statistical significance is indicated based on model p-values; * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Table 10
Use of violence by treatment arm.

	(1) Mother PP	(2) Mother NP	(3) Father PP	(4) Father NP	(5) Violent Ever	(6) Violent Daily
Core	0.220***	−0.461***	0.156***	−0.272***	−0.0129	−0.0160*
	(0.0666)	(0.0528)	(0.0492)	(0.0672)	(0.0303)	(0.00813)
Core Plus	0.403**	−0.342***	0.210**	−0.189*	0.0590	0.00338
	(0.142)	(0.102)	(0.0836)	(0.100)	(0.0363)	(0.0143)
Controls	Yes	Yes	Yes	Yes	Yes	Yes
Observations	1592	1589	1518	1514	1517	1517
F-T1vsT2	1.499	1.157	0.464	0.450	11.39	2.095
F-pval	0.240	0.299	0.506	0.512	0.00416	0.168

Notes: outcome variable is in standard deviations of reference population. Control variables (not shown) include baseline value of the outcome variable, as well as all baseline child and caregiver controls (see Table 1). *Int* stands for interactions; *PP* stands for positive parenting; *NP* stands for negative parenting. Standard errors in parentheses clustered at the ward level; statistical significance is indicated based on model p-values; * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Core intervention had a substantial effect on parenting practices, but little-to-none on the use of violence to resolve conflict between partners. We conclude that the Plus package of interventions did not succeed in reducing violence as intended. Our data are silent on the potential for violence reduction to improve early child development indicators; although the Core intervention did reduce the use of disciplinary parenting practices, we cannot separate out the effect of this from other changes induced by the programme.

6.1.4. Migration

The migration of primary caregivers for work, with or without their children, could have substantial implications for child development. On the one hand, the migration of a caregiver could result in an increase in family income, which could positively impact development. On the other, such migrations could reduce the amount of time and attention given to the child, either by reducing time budgets within the household or by placing the child in alternative care arrangements where the new caregivers are themselves already overstretched. Recent research from China, where many parents migrate away from rural areas for work, suggests that left-behind children have lower levels of stimulation at home (Yue et al., 2020). Young children whose mothers migrated for work also have reduced cognitive development (Yue et al., 2020), and a greater probability of cognitive delay (Bai, Yang, Wang, & Zhang, 2022).

Like many LMICs, Tanzania has experienced substantial rural–urban migration over the past four decades, with persistent income differentials providing continued incentives to migrate (see, e.g., Aikaeli, Mtui, & Tarp, 2021). The area of our study is no exception: while the nearby town Mbeya and surrounding areas attract a substantial number of migrants, Ocello, Petrucci, Testa, and Vignoli (2015) show that Mbozi is a district with particularly high levels of out-migration. Out-migration of caregivers from our study area could have two particular effects on the children in our study sample: first, it could cause attrition from the sample, and second, it could affect the development of children who are left behind. We will discuss these two concerns separately.

While we do not know the reasons why respondents were not contactable at follow up, we can study the characteristics of panel members and attriters. Using data from Tanzania's Integrated Labour Force Survey (2006), Msigwa and Mbongo (2013) find that household income, skill level and education are positively associated with migration, and that men and married individuals are more likely to migrate than unmarried. Older individuals, and those with larger families, are less likely to migrate. Table 17 shows that children who attrited from the sample were more likely to be first-born (i.e. from smaller families) and have young mothers - characteristics positively associated with migration. We also find, however, that panel members are more likely to have a literate mother, and to be girl children. This pattern suggests that, while it is likely that some of the attrition from our sample is due to caregiver out-migration, it is not the only force at play.

In the absence of data on migration, we control for the propensity to migrate as best we can. Ingelaere, Christiaensen, Weerdt, and Kanbur (2018) highlight the importance of secondary towns for young rural Tanzanians' initial migration decision. Given the great distance between our study site and Tanzania's largest cities (800 km to Dar es Salaam and 900 km to Mwanza), the draw of secondary towns is likely to be crucial to the migration decision of individuals in our sample. In addition to a rich set of household controls covering many of the characteristics Msigwa and Mbongo (2013) identified as important, we also include a set of distance measures in our primary specification: distance from interview location to any road, to the local highway, and to the nearest town.

Our data do show who the panel children are living with. We consider two proxies for left-behind status: children whose primary caregiver is not their mother, and children who live with only one of their parents. At wave 2, 137 children have a primary caregiver who is not their mother (T = 65, C = 74). These children do not have statistically significantly different CREDI scores compared with the rest of the sample (0.019 vs 0.067, p = 0.5911). At wave 2, 104 children live with only one of their parents (in all cases this is the mother; T = 64, C = 41). While the difference in point estimates is larger, there is again no statistically significant difference in CREDI scores between these children and the rest of the sample (-0.019 vs 0.072, p = 0.3551). This exploration gives us confidence that the absence of migration data in our model is not driving the results. It would be helpful if future work investigated this more carefully, as the migration status of parents could have important implications for the design and the effectiveness of caregiver training programmes.

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 and measures of remoteness raises some concerns. While the primary specification in all regressions controls for the baseline levels of both demographic 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 par-

Table 11
Child development: varying sets of controls.

	(1) Overall	(2) Overall	(3) Overall
Treat	0.271***	0.236***	0.0975
	(0.0782)	(0.0793)	(0.0725)
Basic Controls	Yes	Yes	Yes
Wave 0 Targeted Practices	No	Yes	Yes
Wave 1 Targeted Practices	No	No	Yes
Observations	1502	1502	1502

Notes: regressions include varying sets of controls. Basic: child and caregiver sex, child age, first child dummy, household SES, mother and father literacy, young mother dummy, distance to a major road, distance to the primary highway, and distance to the nearest town. Practices: play materials, mother and father interaction activities, mother and father positive and negative parenting. For comparability, all regressions are done with the same set of observations. Standard errors in parentheses clustered at the ward level; statistical significance is indicated based on model p-values; * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

enting style at baseline is not driving the findings. This also raises the question: 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. If measurable changes in these characteristics are driving the change in child development scores, adding these as controls should reduce the estimated treatment effect.

Table 11 reports the treatment effect estimated from Eq. 1 using these three different sets of controls. First, in Column (1), only demographic controls are included; Column (2) includes both demographic and home and caregiver controls from baseline (our primary specification), while Column (3) also includes home and caregiver controls from wave 2. The treatment effects estimated across the first two specification are very similar, although the point estimate is slightly higher when fewer controls are included.[7] Once the wave 2 values of home and caregiver indices have been controlled for (Column (3)), 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 similar point estimate, compared with Column (1).

6.2.2. Attrition

The overview of attrition in Section 3.5 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 three ways: first, by re-weighting the sample; second through a bounding exercise; and last by re-estimating the main results using propensity score matching.

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 Eq. 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 Eq. 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 (i.e. most likely to attrit).

Table 12 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.

We next 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 Eq. 1 with no controls, and calculate bounds for that estimator. This gives some sense of the scope for attrition to bias the results.

Table 13 reports treatment effect 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.

Finally, we re-estimate our results using propensity score matching. We carry out a one-to-one matching without replacement using our full set of baseline controls, implemented in Stata using PSMATCH2 (Leuven & Sianesi, 2003). To explore the sensitiv-

[7] Note that the estimated treatment effect in Column (2) does not exactly match the main results in Table 3. Since our interest here is comparing the estimates under different sets of controls, we have restricted the sample in each column to those observations with complete data for the maximum set of control variables.

Table 12
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	0.281***	0.257***	0.366***	0.335***	0.315***	−0.406***	0.192***	−0.230***
	(0.0751)	(0.0571)	(0.0645)	(0.0757)	(0.0956)	(0.0587)	(0.0586)	(0.0567)
Controls	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Observations	1638	1630	1636	1636	1592	1589	1518	1514

Notes: outcome variables are in standard deviations units of the reference population (CREDI) or the control group (all others). Control variables (not shown) include baseline value of the outcome variable, as well as all baseline child and caregiver controls (see Table 1). *Int* stands for interactions; *PP* stands for positive parenting; *NP* stands for negative parenting. Regressions are weighted using inverse-probability weights. Standard errors in parentheses clustered at the ward level; statistical significance is indicated based on model p-values, * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Table 13
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***	0.278***	0.473***	0.312***	0.359***	−0.346***	0.175***	−0.210***
	(0.0472)	(0.0492)	(0.0539)	(0.0556)	(0.0555)	(0.0501)	(0.0553)	(0.0497)
Lower Bound of Treat	0.271	0.278	0.425	0.285	0.345	−0.359	0.165	−0.359
Upper Bound of Treat	0.311	0.288	0.473	0.320	0.378	−0.193	0.189	−0.192
Constant	−0.0857**	0.00275	−0.00242	−0.00982	0.00397	0.00424	−0.000980	0.00252
	(0.0335)	(0.0351)	(0.0384)	(0.0396)	(0.0394)	(0.0356)	(0.0390)	(0.0349)
Observations	1705	1699	1705	1705	1656	1653	1568	1566

Notes: Lower Bound of Treat and *Upper Bound of Treat* are the Lee Bounds on the treatment effect estimate in the first row. Outcome variables are in standard deviations units of the reference population (CREDI) or the control group (all others). No additional control variables are included. *Int* stands for interactions; *PP* stands for positive parenting; *NP* stands for negative parenting. Model standard errors (not clustered) in parentheses; * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Table 14
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?

Table 15
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?

Table 16
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?

ity of our results to quality of the match, we replicate the matching with various amounts of trimming: this process imposes common support by dropping a given percentage of treated observations where the density of controls is least. Trimming in this way should ensure a better match between treated and control units, at least with respect to the characteristics that predict treatment, although it comes at a cost of reducing the effective sample size. We vary the trim from 0% (809 matched pairs) to 50% (405 matched pairs), and re-estimate our primary treatment effect using the resulting subsample. The results, which are shown in Appendix Table 22, are very similar to our main findings.

While none 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 a range of attrition-related concerns.

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)		
Dist to highway as crow flies	1683	14.367	548	13.069	1.298***	0.134
		(0.235)		(0.422)		
Dist to any road as crow flies	1683	1.600	548	1.422	0.178*	0.092
		(0.047)		(0.083)		
Dist to nearest town as crow flies	1683	22.107	548	19.834	2.273***	0.261
		(0.202)		(0.415)		
Overall	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. *Int* stands for interactions; *PP* stands for positive parenting; *NP* stands for negative parenting.

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 0.29 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 £32-£77 per standard deviation increase in child development (details of these calculations, and cost estimates, are in Appendix C).

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

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

Variable	(1) Control N	(1) Control Mean/SE	(2) Treat 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)		
Dist to highway as crow flies	284	16.528	264	9.348	7.180***	0.727
		(0.651)		(0.420)		
Dist to any road as crow flies	284	1.207	264	1.653	−0.446***	−0.229
		(0.106)		(0.129)		
Dist to nearest town as crow flies	284	20.606	264	19.004	1.601*	0.165
		(0.710)		(0.393)		
Overall	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. *Int* stands for interactions; *PP* stands for positive parenting; *NP* stands for negative parenting.

Table 19
Correlation of CREDI, home and caregiver variables at wave 1.

Variables	CREDI	Kinds of toys	Mother Interaction	Father Interaction	Mother PP	Mother NP	Father PP	Father NP
CREDI	1.0000							
Kinds of toys	0.1611	1.0000						
	(0.0000)							
Mother Interaction	0.1171	0.2727	1.0000					
	(0.0000)	(0.0000)						
Father Interaction	0.0951	0.2596	0.4992	1.0000				
	(0.0000)	(0.0000)	(0.0000)					
Mother PP	0.0764	0.1536	0.2821	0.1865	1.0000			
	(0.0003)	(0.0000)	(0.0000)	(0.0000)				
Mother NP	0.0597	0.1293	0.1207	0.0717	0.1955	1.0000		
	(0.0046)	(0.0000)	(0.0000)	(0.0007)	(0.0000)			
Father PP	0.1005	0.1472	0.2236	0.3027	0.7475	0.1567	1.0000	
	(0.0000)	(0.0000)	(0.0000)	(0.0000)	(0.0000)	(0.0000)		
Father NP	0.0456	0.1301	0.0872	0.1289	0.1307	0.6484	0.1997	1.0000
	(0.0305)	(0.0000)	(0.0000)	(0.0000)	(0.0000)	(0.0000)	(0.0000)	

Notes: obs = 2,254. *PP* stands for positive parenting; *NP* stands for negative parenting.

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

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

Variable	(1) Did not attend N	Mean/SE	(2) Attended N	Mean/SE	T-test Difference (1)-(2)	Normalized difference (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)		
Dist to highway as crow flies	419	8.953	414	10.303	−1.350***	−0.205
		(0.335)		(0.306)		
Dist to any road as crow flies	419	1.462	414	1.510	−0.048	−0.027
		(0.087)		(0.089)		
Dist to nearest town as crow flies	419	19.269	414	21.099	−1.830***	−0.320
		(0.254)		(0.298)		
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. *Int* stands for interactions; *PP* stands for positive parenting; *NP* stands for negative parenting.

Table 21
Multiple hypothesis corrections.

	Model	Bonferroni		Cluster			Cluster and stratify		
		10%	5%	Re-sample	Diff/model	Romano-Wolf	Re-sample	Diff/model	Romano-Wolf
Motor	0.0255			0.0925	2.6250	0.1181	0.0275	0.0776	0.0570
Cognitive	0.0038	+	*	0.0269	6.1004	0.1010	0.0072	0.9003	0.0463
Language	0.0022	+	*	0.0188	7.5424	0.0899	0.0140	5.3608	0.0445
Socio-emotional	0.0066	+		0.0303	3.6192	0.1181	0.0123	0.8749	0.0570
Toys	0.0007	+	*	0.0044	5.4349	0.0618	0.0106	14.501	0.0433
Mother Int	0.0000	+	*	0.0031	149.85	0.0178	0.0013	62.252	0.0111
Father Int	0.0004	+	*	0.0060	13.112	0.0540	0.0056	12.170	0.0396
Mother PP	0.0058	+		0.0515	7.8083	0.1181	0.0221	2.7795	0.0570
Mother NP	0.0000	+	*	0.0013	341.16	0.0097	0.0012	314.81	0.0060
Father PP	0.0070	+		0.0298	3.2762	0.1181	0.0196	1.8123	0.0570
Father NP	0.0009	+	*	0.0090	9.2731	0.0662	0.0108	11.327	0.0433

Notes: Diff/model is the difference between the re-sample and model p-values, divided by the model p-value. Cluster is by ward; stratification (last 3 columns) is by child sex, caregiver sex and treatment. *Int* stands for interactions; *PP* stands for positive parenting; *NP* stands for negative parenting.

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 present analysis 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 care-

Table 22
Overall CREDI score: Propensity score matching.

	(1) Trim 0	(2) Trim 5	(3) Trim 15	(4) Trim 25	(5) Trim 50
Treat	0.281***	0.278***	0.279***	0.277***	0.320***
	(0.0762)	(0.0735)	(0.0753)	(0.0732)	(0.0693)
Controls	Yes	Yes	Yes	Yes	Yes
Observations	1604	1525	1364	1203	803

Notes: outcome variable is overall CREDI score, in standard deviations units of the reference population. Control variables (not shown) include baseline value of the outcome variable, as well as all baseline child and caregiver controls (see Table 1). Standard errors in parentheses clustered at the ward level; statistical significance is indicated based on model p-values, * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

givers 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 practice-led caregiver-focused 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 receptive to change.

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? In a recent review, Jeong, Pitchik, and Fink (2021) highlight a scarcity of evidence on the medium to long term effectiveness of parenting interventions in LMICs. The available evidence suggests that effects fade relatively quickly, with few studies finding effects lasting beyond 1–3 years. 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 effectiveness, and ensure that adaptations of the intervention to new contexts 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.

Data availability

The authors do not have permission to share data.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgement

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, Damian Clarke, Deborah Cobb-Clark, Thomas Dohmen, David Escamilla-Guerrero, Jakob Hennig and David Jaeger, as well as seminar participants at the University of Essex Institute for Social and Economic Research, the University of Glasgow and the University of St Andrews for helpful comments and discussions. Comments from the editor and two anonymous referees substantially improved the paper. Research assistance by Lauren Davis, Lawrence Ho, Himangshu Kumar and Peter Lowley 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.

Appendix A. Data appendix

A.1. Constructed variables

A.1.1. Home environment

Table 14 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.

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 15 lists the full set of questions. The possible responses were: once; a few times (two or three); frequently (more than 3 times).

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 16. 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.

A.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. The attrited observations are also *less likely* to live in remote areas: while the size of these differences is not that great, it is a consistent trend across all our measures of remoteness.

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. In terms of remoteness, attriters in the treatment group resided (at wave 1) closer to the main highway, but farther from the nearest road. They were also slightly closer to a town.

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, although those who attended lived in slightly more remote locations.

Appendix B. Extended results

B.1. Multiple hypothesis testing

We apply multiple hypothesis corrections to the secondary treatment effect estimates presented in Tables 4 & 5.[8] The tables report the treatment effect of the intervention on four subdomains of child development (Table 4) and seven home and caregiver outcomes (Table 5), for a total of 11 separate hypothesis tests. We carry out two multiple hypothesis correction procedures: the Bonferroni correction and the Romano-Wolf correction. We include the Bonferroni correction, which is easy to compute but less appropriate for our purposes than Romano-Wolf, as a check against the more sophisticated method.

The Bonferroni correction is an early approach to correcting traditional significance tests for multiple hypotheses. This method adjusts the significance threshold for rejecting the null hypothesis by the number of hypotheses being tested – specifically, by dividing the significance level by the number of tests. This method is very conservative: the adjustment gives confidence that statistically significant results are not spurious findings due to repeated testing (it has strong control), but it carries an elevated risk of falsely dismissing findings as statistically insignificant (failing to reject a false null hypothesis). A good overview of this approach can be found in Clarke et al. (2020). In our case, where we are testing 11 hypotheses at once, a single-test significance threshold of 10% (which would imply rejecting tests with $p < 0.10$ without correction), would correspond to a corrected significance threshold of $(10\%/11 = 0.9\%)$, for a rejection condition of $p < 0.009$. A single-test significance threshold of 5% would imply rejecting tests with $p < (0.05/11) = 0.0045$.

The Romano-Wolf correction, introduced by Romano and Wolf (2005) and Romano and Wolf (2005), 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 rejecting at least one true null hypothesis amongst this set. The Romano-Wolf correction method uses re-sampling to estimate the correlation between test statistics. This allows for a less conservative adjust-

[8] The authors are grateful to Damian Clarke for assistance in understanding and clarifying this section.

ment compared with methods, such as Bonferroni, that allow for any correlation structure by taking the worst-case-scenario (Clarke et al., 2020).

The fact that Romano-Wolf accounts for correlation within the data makes it well-suited to our context, where our outcomes are all related to each other. Its greater power is also highly desirable in impact evaluation applications, where sample sizes are generally modest. The real-world cost of failing to reject a false null hypothesis (e.g. failing to find a statistically significant difference, when that difference does exist) is also potentially large in such settings. A downside of Romano-Wolf is that it is considerably more complicated to implement. User-friendly packages now exist, however getting a reliable correction is not always straightforward. Because the correction relies on re-sampling and bootstrap methods, Romano-Wolf is only reliable in practice when the re-sample procedure generates a sufficiently accurate approximation of the original (model) p-values. This can be hard to achieve with more complex models, e.g. with clustered standard errors. Without an accurate re-sample, the correction is not meaningful: it is therefore useful to present the re-sampled p-values along with the model and corrected p-values.

We apply the rwolf2 Romano-Wolf correction implemented in Stata by Clarke (2021) (an update to the earlier rwolf, by Clarke et al. (2020)). Within the rwolf2 command, we simultaneously estimate the 11 equations presented in Tables 4 & 5, with standard errors clustered at the ward level and 10,000 repetitions. To ensure that the resampling procedure, and the bootstrap based on this, accounts for this clustering, we include the cluster(ward) and idcluster(newclust) options. As we struggled to replicate the model p-values with this method (re-sampled p-values often many times larger than the model), we sought to improve the re-sample procedure by adding stratification on key variables: child sex, caregiver sex and treatment status. While the re-sampled p-values are still somewhat larger than the model (which implies the correction will be overly conservative), the correspondence with the model is much better. We therefore present this specification as our primary Romano-Wolf correction.

A summary of our multiple hypothesis corrections is given in Table 21. The first column presents the model (unadjusted) p-values for the treatment effect estimate for each regression in Tables 4 & 5. These unadjusted p-values suggest that all 11 treatment effect estimates are highly significant, with motor development significant at the 5% level and all others at the 1% level. The next two columns indicate whether or not each treatment effect would remain statistically significant after a Bonferroni correction. While all but motor development remain significant at the 10% level, only 7 of the 11 estimates remain significant at the 5% level.

The last six columns present the Romano-Wolf correction, first with only clustering, and then with clustering and stratification (by child sex, caregiver sex and treatment). Without stratification, the re-sampling procedure struggles to replicate the model p-values: the difference in p-values range from 2.6 to more than 10 times the model value (in some cases much more, but only when the model p-value is almost 0). The corrected p-values are, as expected, larger still. While some treatment effect estimates remain significant despite this, nearly half of them do not meet the 10% significance threshold. The Romano-Wolf correction with clustering and stratification performs better: although the p-values are still considerably larger than the model, they are in the correct range (again, with the exception of very small model values). The correction procedure inflates these re-sampled p-values; however, all the estimates remain significant at the 10% level, and only 4 estimates fall (just) short of the 5% significance threshold.

Based on this, we conclude that our secondary treatment effect results are reasonably robust to multiple hypothesis testing corrections. The Romano-Wolf adjustment based on our most accurate re-sampling procedure does not overturn any conclusions based on uncorrected p-values, although it does dial back the strength of our confidence in their statistical significance.

B.2. Propensity score matching

Table 22 presents the results from the propensity score matching exercise described in Section 6.2.2. Each column shows the results for a different amount of trim (removing treated observations where the support is weakest), from 0% to 50%.

Appendix C. 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 and 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.

C.1. Costs

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.[9] Of this, £259,499 are direct project costs (e.g. material development, consumables, community facilitator stipend, advocacy, technical support - but *excluding* salaries, monitoring and evaluation, overheads, capacity development and capital costs); £97,948 (38%) of which are attributed specifically to the Core caregiver-focused 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 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

[9] 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 23
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.

Table 24
Cost per standard deviation improvement: average and marginal costs.

Cost approach		ITT 0.29	ToT 0.55
AC1	£179.08	£617.51	£325.60
AC2: ITT	£114.19	£393.76	
AC2: ToT	£228.37		£415.22
MC1: ITT	£21.21	£73.14	
MC1: ToT	£42.42		£77.13
MC2: ITT	£9.30	£32.07	
MC2: ToT	£18.60		£32.84

Notes: AC is average cost and MC is marginal cost. Four different costing approaches are described in the text and calculated in Table 23. 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 Treatment 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.

therefore double the number who attended at least one session (2 x 6,118 = 12,236).

C.2. Cost effectiveness

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 estimated average treatment effect on the treated, 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.[10]

Table 23 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 focused 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 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 24. 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 24 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 £325. 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 £73 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 £32 per standard deviation increase in early child development.[11]

References

Aikaeli, J., Mtui, J., & Tarp, F. (2021). Rural-Urban Migration, Urbanisation and Unemployment: The Case of Tanzania Mainland. African Journal of Economic Review, *IX*, 87–108.

Alderman, H., Friedman, J., Ganga, P., Kak, M., & Rubio-Codina, M. (2021). Assessing the performance of the Caregiver Reported Early Development Instruments (CREDI) in rural India. *Annals of the New York Academy of Sciences, 1492*, 58–72.

Andrew, A., Attanasio, O., Fitzsimons, E., Grantham-McGregor, S., Meghir, C., & Rubio-Codina, M. (2018). Impacts 2 years after a scalable early childhood

[10] 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.

[11] 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.

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.
Bai, Y., Yang, N., Wang, L., & Zhang, S. (2022). The impacts of maternal migration on the cognitive development of preschool-aged children left behind in rural China. *World Development, 158* 106007.
Black, M.M., S.P. Walker, L.C.H. Fernald, et 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, et al. (2017): Nurturing care: promoting early childhood development, The Lancet, 389, 91 – 102.
Clarke, D. (2021): rwolf2 Implementation and Flexible Syntax, Unpublished.
Clarke, D., Romano, J. P., & Wolf, M. (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, Elsevier, vol. 4 of Handbook of Labor Economics, 1315–1486.
Desiere, S., Vellema, W., & D'Haese, M. (2015). A validity assessment of the Progress out of Poverty Index (PPI). *Evaluation and Program Planning, 49*, 10–18.
Doyle, O., Harmon, C. P., Heckman, J. J., & Tremblay, R. E. (2009). Investing in early human development: Timing and economic efficiency. *Economics & Human Biology, 7*, 1–6.
Emmers, D., Jiang, Q., Xue, H., Zhang, Y., Zhang, Y., Zhao, Y., Liu, B., Dill, S.-E., Qian, Y., Warrinnier, N., Johnstone, H., Cai, J., Wang, X., Wang, L., Luo, R., Li, G., Xu, J., Liu, M., Huang, Y., Shan, W., Li, Z., Zhang, Y., Sylvia, S., Ma, Y., Medina, A., & Rozelle, S. (2021). Early childhood development and parental training interventions in rural China: a systematic review and meta-analysis. *BMJ Global Health, 6* e005578.
Grantham-McGregor, S., Adya, A., Attanasio, O., Augsburg, B., Behrman, J., Caeyers, B., Day, M., Jervis, P., Kochar, R., Makkar, P., Meghir, C., Phimister, A., Rubio-Codina, M., & Vats, K. (2020). Group Sessions or Home Visits for Early Childhood Development in India: A Cluster RCT. *Pediatrics, 146* e2020002725.
Ingelaere, B., Christiaensen, L., Weerdt, J. D., & Kanbur, R. (2018). Why secondary towns can be important for poverty reduction - A migrant perspective. *World Development, 105*, 273–282.
Jeong, J., Adhia, A., Bhatia, A., McCoy, D. C., & Yousafzai, A. K. (2020). Intimate Partner Violence, Maternal and Paternal Parenting, and Early Child Development. *Pediatrics, 145* e20192955.
Jeong, J., Pitchik, H. O., & Fink, G. (2021). Short-term, medium-term and long-term effects of early parenting interventions in low- and middle-income countries: a systematic review. *BMJ Global Health, 6* e004067.
Lee, D. S. (2009). Training, Wages, and Sample Selection: Estimating Sharp Bounds on Treatment Effects. *Review of Economic Studies, 76*, 1071–1102.
Jeong, J., E.E. Franchett, C.V.R. d. Oliveira, K. Rehmani, & A.K. Yousafzai (2021a): Parenting interventions to promote early child development in the first three years of life: A global systematic review and meta-analysis, PLoS Medicine, 18, e1003602.
Leuven, E. & B. Sianesi (2003): PSMATCH2: Stata module to perform full Mahalanobis and propensity score matching, common support graphing, and covariate imbalance testing, Stata module.
Li, Y., Tang, L., Bai, Y., Zhao, S., & Shi, Y. (2020). Reliability and validity of the Caregiver Reported Early Development Instruments (CREDI) in impoverished regions of China. *BMC Pediatrics, 20*, 475.
Lu, C., Black, M. M., & Richter, L. M. (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.
Luoto, J. E., Garcia, I. L., Aboud, F. E., Singla, D. R., Fernald, L. C. H., Pitchik, H. O., Saya, U. Y., Otieno, R., & Alu, E. (2021). Group-based parenting interventions to promote child development in rural Kenya: a multi-arm, cluster-randomised community effectiveness trial. *The Lancet Global Health, 9*, e309–e319.
McCoy, D.C., Fink, G. Waldman M. (2018a): CREDI Data Management & Scoring Manual, User guide.
McCoy, D. C., Seiden, J., Waldman, M., & Fink, G. (2021). Measuring early childhood development: considerations and evidence regarding the Caregiver Reported Early Development Instruments. *Annals of the New York Academy of Sciences, 1492*, 3–10.
McCoy, D. C., Sudfeld, C. R., Bellinger, D. C., Muhihi, A., Ashery, G., Weary, T. E., Fawzi, W., & Fink, G. (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., Waldman, M., CREDIFieldTeam1 & Fink, G. (2018). Measuring early childhood development at a global scale: Evidence from the Caregiver-Reported Early Development Instruments. *Early Childhood Research Quarterly, 45*, 58–68.
Msigwa, R. E., & Mbongo, J. E. (2013). Determinants of Internal migration in Tanzania. *Journal of Economics and Sustainable Development, 4*, 28–35.
Munoz-Chereau, B., Ang, L., Dockrell, J., Outhwaite, L., & Heffernan, C. (2021). Measuring early child development across low and middle-income countries: A systematic review. *Journal of Early Childhood Research, 19*, 443–470.
Norman, R. E., Byambaa, M., De, R., Butchart, A., Scott, J., & Vos, T. (2012). The Long-Term Health Consequences of Child Physical Abuse, Emotional Abuse, and Neglect: A Systematic Review and Meta-Analysis. *PLoS Medicine, 9* e1001349.
Ocello, C., Petrucci, A., Testa, M. R., & Vignoli, D. (2015). Environmental aspects of internal migration in Tanzania. *Population and Environment, 37*, 99–108.
Oliveira, C.V.R. d., Sudfeld, C.R., Muhihi, A., McCoy, D.C., Fawzi, W.W., Masanja, H., Yousafzai A.K. (2022): Association of Exposure to Intimate Partner Violence With Maternal Depressive Symptoms and Early Childhood Socioemotional Development Among Mothers and Children in Rural Tanzania, JAMA Network Open, 5, e2248836.
Ozler, B., Fernald, L. C. H., Kariger, P., McConnell, C., Neuman, M., & Fraga, E. (2018). Combining pre-school teacher training with parenting education: A cluster-randomized controlled trial. *Journal of Development Economics, 133*, 448–467.
Panter-Brick, C., Burgess, A., Eggerman, M., McAllister, F., Pruett, K., & Leckman, J. F. (2014). Practitioner Review: Engaging fathers – recommendations for a game change in parenting interventions based on a systematic review of the global evidence. *Journal of Child Psychology and Psychiatry, 55*, 1187–1212.
Richter, L.M., Daelmans, B., Lombardi, J., et 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., & Wolf, M. (2005). Exact and Approximate Stepdown Methods for Multiple Hypothesis Testing. *Journal of the American Statistical Association, 100*, 94–108.
Romano, J. P., & Wolf, M. (2005). Stepwise Multiple Testing as Formalized Data Snooping. *Econometrica, 73*, 1237–1282.
Russell, A.L., Hentschel, E., Fulcher, I., Rav, M.S., Abdulkarim, G., Abdalla, O., Said, S., Khamis, H., Hedt-Gauthier, B., Wilson K. (2022): Caregiver parenting practices, dietary diversity knowledge, and association with early childhood development outcomes among children aged 18-29 months in Zanzibar, Tanzania: a cross-sectional survey, BMC Public Health, 22, 762.
Sabol, T. J., McCoy, D., Gonzalez, K., Miratrix, L., Hedges, L., Spybrook, J. K., & Weiland, C. (2022). Exploring treatment impact heterogeneity across sites: Challenges and opportunities for early childhood researchers. *Early Childhood Research Quarterly, 58*, 14–26.
Schreiner, M. (2016): PPI for Tanzania 2011, Progress out of Poverty.
Sylvia, S., Luo, R., Zhong, J., Dill, S.-E., Medina, A., & Rozelle, S. (2022). Passive versus active service delivery: Comparing the effects of two parenting interventions on early cognitive development in rural China. *World Development, 149* 105686.
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*. Dar es Salaam, Tanzania: UNICEF. Tech. rep.
Yue, A., Bai, Y., Shi, Y., Luo, R., Rozelle, S., Medina, A., & Sylvia, S. (2020). Parental Migration and Early Childhood Development in Rural China. *Demography, 57*, 403–422.