Impacts of a Large-Scale Parenting Program: Experimental Evidence from Chile[1]

Pedro Carneiro
University College London, IFS, CeMMAP, FAIR-NHH

Emanuela Galasso
World Bank

Italo Lopez Garcia
Center of Economic and Social Research
University of Southern California

Paula Bedregal
Pontificia Universidad Católica, Chile

Miguel Cordero
Universidad de Chile

December 2022

Abstract: This paper presents results from a large-scale experimental evaluation of the impacts on parents and children of a national parenting program in Chile. The program is low cost: it lasts only for six to eight weeks, and it is administered to groups (of 8 to 12 parents) rather than individuals. It is implemented by the national health system, taking advantage of its existing physical infrastructure, and the deep staff knowledge about the constraints faced by parents and children in each location. We find that children whose parents are offered the opportunity to participate in this program experience increases in their vocabulary and socio-emotional development scores of 0.1 standard deviations (SD), which are mirrored by improvements in parenting behaviors and parenting beliefs of the same magnitude. These impacts are observed almost three years after the intervention ends, far outlasting its duration.

JEL No: H43, I10, I20, I38
Keywords: parenting, early childhood development

[1] We thank participants in several seminars and conferences for comments. This project was funded by the World Bank Research Budget, the Knowledge for Change program and the Strategic Impact Evaluation Fund. Pedro Carneiro gratefully acknowledges the support of the ESRC for CEMMAP (ES/P008909/1) and the European Research Council through grant ERC-2015-CoG- 692349. Miguel Cordero acknowledges to the Scholarship Formation of the Advanced Human Capital Programme, Becas Chile. A special thank goes to Lucia Vergara, Cecilia Moraga and Felipe Arriet at the Ministry of Health for their support of the evaluation, to Veronica Silva for her support in setting up the evaluation, and to Veronica Mingo and the staff of *Juguemos con Nuestros Hijos* for the adaptation of the curriculum for the intensive arm of the evaluation. We are grateful to Ruben Poblete Cazenave and Nicolas Libuy Rios for able research assistance. Finally, our greatest thanks go to all parents and children who participated in this study and to the local facilitators and their primary care teams who helped to make this study happen. All errors and omissions are own.

1 Introduction

A large body of work across disciplines has documented the importance that early experiences and environments play in shaping child health, cognitive and socio-emotional development. Children living in poverty experience greater levels of environmental and psychosocial stressors that lead to diverging trajectories from very early ages (Heckman and Mosso 2014). There is compelling evidence that parenting behaviors and the quality of the home environment play a critical role during the sensitive periods of early childhood (Cunha and Heckman 2008, Todd and Wolpin 2007, Cunha, Heckman, and Schennach 2010).

Despite the widespread consensus that early childhood development (ECD) interventions from birth to five years of age have the potential to positively impact adult outcomes, most of the existing evidence stems from small-scale pilots testing intensive interventions working directly with children in child-care centers (Campbell and Ramey 1994, Heckman et al. 2010), or with parents and children in home visiting programs (Gertler et al. 2014, Heckman et al. 2017). Yet such intensive programs are often too expensive to implement at scale, and little is known about the effectiveness of ECD interventions that focus on parents which are implemented at a national level, especially in low- and middle-income countries (LMICs).[2]

In this paper, we present experimental medium-term impacts of *Nadie es Perfecto* (NEP), a low-cost group-based parenting program delivered nationwide by the national health system in Chile. Outcomes are measured roughly three years after the end of the program completion. Two different versions of this program are offered (randomly) to potential participants: NEP-Basico (NEP-B) and NEP-Intensivo (NEP-I). NEP-B consists of eight weekly group parenting sessions, while NEP-I supplements these with two additional sessions where caregivers can interact with their children and receive feedback on their practices.

We show that this program had sustained impacts on children's developmental outcomes, as well as on parenting behaviors and beliefs, three years after caregivers completed participation. Children whose parents were offered participation in NEP-I have cognitive scores which are 0.1 standard deviations (SD) above children whose parents were not offered to participate in NEP-I. We document similar impacts for socioemotional outcomes. Impacts of NEP-I on standardized indices of parents' parenting behaviors and beliefs are of about the same magnitude as those reported for child outcomes. NEP-B has smaller impacts on children and parents than NEP-I.[3]

NEP is worthwhile studying for three main reasons. First, NEP is not a pilot intervention implemented in a small local area. On the contrary, NEP-B, the benchmark version of NEP, is a scaled-up, publicly provided parenting program for families that are beneficiaries of the primary health care system and therefore are generally poor. The program is fully integrated with other health, education, and welfare programs and is delivered through health centers across the whole

[2] A growing body of research from low- and middle-income countries has shown that a broader set of ECD parenting interventions are effective to improve parenting behaviors and child development outcomes, at least in the short term (Aboud and Yousafzai 2015, Jeong et al. 2021) , but evidence on longer term outcomes is much sparser (Jeong, Pitchik, and Fink 2021).

[3] The impacts described in this paragraph are Intent-to-Treat (ITT) estimates that do not adjust for less than full take-up of the program by those offered a chance to enroll in it. The impacts of having participated in the program on these same cognitive and socioemotional outcomes, estimated by instrumental variables (IV), are about 0.4 SD.

country. Even NEP-I, which was developed through a joint venture of researchers and the Ministry of Health, is only mildly more intensive than NEP-B, and delivered using the same human and physical infrastructure. Facilitators of NEP sessions are professional staff from local health centers: nurses, educators, psychologists, and social workers. They have a deep knowledge of the target population for NEP in the locations they work in, and they have strong ties to the community. The combination of these traits with 32 hours of training for the standard NEP-B program, and 16 additional hours of training for the two extra sessions included in the NEP-I program, appears to have made them very effective group facilitators.

Second, our team had the opportunity to conduct a large-scale experimental evaluation of this national program, enabling us to construct credible estimates of program impacts. Furthermore, the close cooperation between our research team and civil servants from the Ministries of Health and Social Planning enabled us not only to implement a rigorous research design, but also to collect high quality household and child data at two points in time. The combination of an experimental design with detailed survey data is unusual for government programs of this scale.

The experimental research design was possible because capacity constraints prevented all potentially eligible families to be served simultaneously, giving researchers the opportunity to randomize the order in which families accessed the program. Since so many families were deemed eligible for the program when its implementation first started, significant time lags occurred between the time the first and the last sets of families in our sample had access the program.

Third, even if delivered at small scale or in a pilot setting, NEP would be an interesting program to study given its innovative features. NEP is a semi-structured parenting program with a behavioral approach based on social learning principles (Bandura 1995, Bandura 1986). It has components of a structured curriculum, but (much more than other standard parenting programs) it also tailors the intervention to the needs and interests of each group. During each group session, the facilitator's role is to promote a group discussion on the topics chosen by participant parents for that session. This mode of delivery allows parents to choose specific topics that are better targeted to their needs and enables them to learn both from the structured components of the curriculum and from each other's experiences. Moreover, it is likely to foster the creation of networks of parents typically residing in the same or close by neighborhoods, which may become important sources of support for parents even after the end of the interventions. These elements combined increase the chances that impacts on parenting behaviors and children's development are sustained in the longer-term.

Notably, unlike most evaluations of parenting programs, we are able to document impacts on children and parents almost three years after parents participate in the program. Given the short duration of this intervention, the impacts we report vastly outlast the duration of the program, finding that is remarkably rare among the parenting programs evaluated elsewhere.

This study contributes to the large literature of interventions focused on parenting behavior change with effects that are sustained well beyond the end of the program. Most of the evidence on long-term impacts of high-quality early childhood interventions from both developed and low- and middle-income countries (LMICs) equates high-quality to high cost-high intensity and duration. ECD programs for which we have reported long-term outcomes, such as the Abecedarian or the HighScope Perry Preschool Program, are programs working directly with children in childcare centers many hours a day, effectively bypassing the role of parenting (Campbell and Ramey 1994, Heckman et al. 2010). There also exists evidence that some high-intensity individual home visiting

parenting interventions have sustained early impacts in the medium and long term. Individual home visits provide an opportunity to tailor activities to individual circumstances to overcome personal barriers to behavioral change, as well as engage with both caregivers and the children with a weekly, biweekly, or monthly contact that lasts between 18 months to 4 years. Home visiting programs from developed countries that have achieved sustained impacts on child development in the medium-term include the Nurse Family Partnership (NFP) in the US (Olds et al. 2004, Heckman et al. 2017) and the Preparing for Life (PFL) in Ireland (Doyle 2020). In LMIC settings, to our knowledge only two home visiting programs have achieved sustained impacts that outlast the program duration. One is the seminal Jamaica Study, a rare instance of a program targeted to disadvantaged children that generated large sustained impacts in adult outcomes in the long run (Grantham-McGregor et al. 1991, Gertler et al. 2014). The second is the Lady Health Worker (LHW) program in Pakistan, a two-year program with impacts two years after the end of the intervention that were still positive but of much reduced magnitude relative to early impacts (Yousafzai et al. 2014, Yousafzai et al. 2016). Recent home visiting programs that build on the Jamaica Study have achieved positive short-term impacts yet substantially smaller in magnitude than those achieved in the seminal program (Araujo et al. 2021, Grantham-McGregor et al. 2020, Attanasio et al. 2014, Sylvia et al. 2020). Moreover, the only study that included a longer-term follow up finds that early impacts completely faded out two years later (Andrew et al. 2018).This is in line with a recent systematic review of parenting interventions in LMICs, most of them home visiting programs, which shows that early impacts tend to fade-out over time (Jeong, Pitchik, and Fink 2021). These results demonstrate that, at the very least, evidence of impacts of home visiting programs is mixed and highly heterogeneous, which is consistent with findings of systematic reviews of parenting interventions worldwide (Jeong et al. 2021). Coupled with their high implementation costs, these mixed results make the individual home visiting model less attractive to implement at scale in most countries.

Group-based parenting interventions do not offer the same personalized attention but provide social support and peer-to-peer learning and are potentially more cost-effective. There is increasing evidence from randomized control trials from LMIC settings that ECD group-based parenting interventions that engage both caregivers and children are at least as effective as individual home visit interventions to improve parental behaviors and child development outcomes, at least in the short-term (Aboud and Yousafzai 2015, Grantham-McGregor et al. 2020, Luoto et al. 2021). However, evidence on the ability of these programs from LMICs to sustain early impacts over time is extremely limited.[4] In developed countries, different meta-analyses find that group-based parenting programs focusing on parents are only effective in improving child socio-emotional and behavioral outcomes (Sanders et al. 2014, Furlong et al. 2012, Barlow et al. 2016). The most widely studied of these programs are the Triple-P Positive Parenting Program (Sanders 2012) and the Incredible Years (Webster-Stratton 2001), both of which are group-based and share the same theoretical basis as NEP. However, all these studies are small scale efficacy trials and only focus on socioemotional outcomes, and, with very few exceptions, they only include short-term evaluations (Kim et al. 2018, Camehl, Spiess, and Hahlweg 2020).[5]

[4] A group-based parenting intervention in Rwanda (Justino et al. 2020) shows sustained impacts three years laterClick or tap here to enter text.. It has a longer duration than NEP (17 weekly sessions instead 8-10 sessions), and provides more in depth support to families through complementary home visits and supporting material/book gift at home.

[5] There exist also interventions that attempt to affect home environments by targeting maternal mental health. One example is (Baranov et al. 2020).

Our results suggest that a group-based parenting intervention with a structured curriculum flexibly tailored to specific parents' interests, can achieve sustained impacts on parenting behaviors and child development at scale, even if it is low cost and low intensity. Its key strength lies in a curriculum that almost exclusively focuses on parents' needs and addresses the barriers and parenting problems that parents face. Importantly, it is the more intensive version of NEP (NEP-I) that combine parents-only sessions with two additional sessions of guided practice with their children that has the highest impact, even though we can never reject the impact is the same across the two interventions. While the two extra sessions including parent-child interactions alone are unlikely to be the reason behind our impacts on children's outcomes, they are likely to have motivated NEP-I facilitators to implement the enhanced curriculum and parents to consolidate knowledge and behavioral change.

This study also contributes to the broader literature on human capital formation centered around the role of parental early investments play in shaping long-term outcomes and compensating for early inequalities (Britto et al. 2017, Aboud and Yousafzai 2015, Heckman and Masterov 2007). A recent body of observational work has documented the role that parental knowledge and beliefs about parenting practices (Cunha, Elo, and Culhane 2013, Attanasio, Cunha, and Jervis 2019, Cunha, Elo, and Culhane 2022), parental beliefs about the importance of parenting (Boneva and Rauh 2018), and parenting styles (Doepke and Zilibotti 2017) play in shaping parenting behaviors. Our NEP study measures and provides experimental evidence on an intervention centered around shifting these underlying determinants, thus contributing to this growing literature.

Relevant to the extrapolation of the results of this study to other settings is the fact that Chile's welfare system is well organized, and staffed with a skilled and motivated workforce, which may have played a role in their success in implementing NEP. The program relies on facilitators with a university degree who, with the appropriate training, are able to deliver a curriculum that is highly flexible to accommodate family needs. One could say that this is a high-quality scale-up, but nevertheless, it is one that is achievable in health systems with similar quality and infrastructure.

More work is needed to test how the curriculum and the intervention can be adapted to other LMIC settings with a lower human resource capacity than Chile. Based on calculations from the Ministry of Health, the cost per child attended per session of the standard NEP-B program is roughly 5-6 times cheaper than a home visit. Our cost-benefit analysis shows that this is a cost-effective program.

The paper proceeds as follows: Section 2 describes the intervention, Section 3 presents the evaluation design, Section 4 describes the data, Section 5 shows our methods and empirical strategy, Section 6 presents the results, Section 7 provides a cost-benefit analysis, Section 8 discusses our findings, and Section 9 concludes.

2 The Intervention

NEP is a parenting intervention operating in the context of a broader early childhood policy platform called *Chile Crece Contigo* (ChCC). The intervention was adapted from a Canadian program, Nobody's Perfect, a long-running group parenting intervention implemented within the public health system in Canada. NEP relies on a semi-structured curriculum that promotes caregivers' knowledge about child development, self-care, positive parenting skills, and the use of non-violent disciplinary strategies, helping caregivers to foster a nurturing home environment.

NEP targets parents with children aged 0 to 5 who are enrolled in the public health system. Potential caregivers are offered participation in the program during regular health check-ups, home visits, or immunization appointments. The intervention can be delivered to all parents who are interested in improving their parenting skills, but it is especially targeted to vulnerable caregivers, such as adolescents, single parents, and geographically or socially isolated households. Parents in these groups can be identified by the health care provider (doctor or nurse) with whom they interact frequently. Households at very high risk (children with severe child developmental delays or disabilities, or high-risk parents with psychiatric problems or intra-household violence) are not considered eligible for NEP and are instead referred to local services with more intensive engagement.

NEP-Basico (NEP-B)

The benchmark version of the program (NEP-B) includes 6 to 8 weekly group sessions with 6-12 caregivers, facilitated by a trained moderator, and based on a curriculum that promotes positive parenting skills to improve cognitive stimulation, to manage child behavior with positive disciplinary strategies, and to improve their parental self-esteem. Each session lasts approximately two hours.

There are several features that distinguish NEP-B from other group-based interventions and are worth highlighting. The first key innovation of the approach lies in a semi-structured curriculum that fosters parental competence by tailoring the intervention to the group's interests and needs. This flexibility is novel and important, allowing parents to choose the specific topics for each session (organized along physical development, cognitive development, behavior, safety, and parental self-care).[6]

A second feature is that NEP-B is based on a model of experiential learning designed for adults (Kolb 2014).[7] That is, during the 2-hour session, the role of the facilitator is to promote a group discussion about the topics chosen by parents for that session. Parents not only acquire new knowledge about their children but also discuss self-care and engage in some introspection about themselves as parents. They are encouraged to learn from each other experiences and discuss the common barriers to enacting new behaviors. Moreover, by fostering the creation of new relationships among parents in the same group, the program has the potential to create social connections outside the group as parents live relatively close to each other. Finally, caregivers are

[6] The topics are covered in five books, which are distributed to participants: 1) Physical development, including topics such as physical growth, health, nutrition and early detection of common illnesses in early years; 2) Mental development, including topics such as cognitive and emotional development, the role of playing and how to stimulate a child according to their age; 3) Behavior, designed as a guide on common behavioral problems and their effective management and resolution using positive disciplinary strategies; 4) Safety and prevention, designed to identify, prevent and manage common risks and accidents at home, including first aid training; 5) Parental and caregiver's self-care, involving activities to improve parental self-image, self-help in the parenting task, the prevention of domestic violence and the promotion of healthy habits strategies for adults. Both caregivers and facilitators are provided with additional materials (stickers with emergency phone numbers on them, promotional posters of NEP for parents, audiovisual, and board games for facilitators).

[7] The training workshop looks at introducing facilitators to the model, learning the goals of the program, and how to use the eligibility criteria to select participants who are more likely to benefit from the program. The main goal is that facilitators learn how to conduct a parenting course from beginning to end using a participant-centered method, implementing approaches for adult education, and following the Experience Learning Cycle, a well-established framework to understand how adults can learn.

provided with a set of materials including five books (one per each topic), where each book discusses common problems and strategies to implement at home, as well as stickers with emergency phone numbers and promotional posters of NEP.

The premise of the intervention is that in order to translate knowledge and beliefs into real behavioral change requires an improvement in parental self-image, so that the adoption of positive parenting practices is sustained through positive reinforcements in the parent-child interactions as well as supported by shared norms within the family and the social network of support (Kagitcibasi et al. 2009).

A third distinctive and key feature of NEP-B is that it combines the selection of qualified staff with high-quality training, which enables this program to be less susceptible to the potentially uneven quality of service delivery at scale (Davis et al. 2017). NEP-B is delivered by facilitators who are local professional staff already working in local health centers (such as nurses, educators, psychologists, and social workers), who have a deep knowledge of the target population participating in the program, given by their close and frequent interactions with potential program beneficiaries through the primary healthcare system. In addition, facilitators are trained on the NEP methodology by a set of master trainers, who are certified by the Canadian Nobody's Perfect Program. The focus of the 32-hour training is on active listening skills, and on facilitating group dynamics with flexibility, without forcing discussion themes, or lecturing parents.[8]

NEP-Intensivo (NEP-I)

The (slightly more) intensive version of NEP (NEP-I) was not part of the set of services originally offered by the Ministry of Health. It was developed as an additional evaluation arm for the study, adopted by the Ministry of Health staff administrating NEP, and delivered at scale during the evaluation period. NEP-I was a collaborative effort between the Ministry of Health and a team of child development experts at Pontificia Universidad Catolica (working on the program *Juguemos con Nuestros Hijos*).

NEP-I adds to the standard group intervention two sessions where children are also present, and where caregivers are given the opportunity to interact with their child and receive personalized feedback on their practices. The two added sessions were also conducted in groups and focused on the importance of age-appropriate responsive play (reading children's cues and providing scaffolding, through practice and discussion videos on responsive parent-child interactions) and on the importance of language and reading (through dialogic reading).

Facilitators assigned to the NEP-I arm received two days of extra training (on top of the training to conduct NEP-B) for these sessions. NEP-I facilitators were free to incorporate these two extra sessions at the end of the regular program, or in between the standard NEP sessions. The rationale for the intensive version is to test the value added of offering opportunities for practical demonstration and skill building, which has been shown elsewhere to be associated with effectiveness in parenting interventions (Engle et al. 2011).

The impact evaluation of NEP started at the early stages of the rolling-out of the program. Between the end of 2009 and the beginning of 2010, more than 1,700 facilitators were trained to deliver the NEP-B program. Since then, it has been fully scaled up at the national level and more than 150,000 families have participated. NEP is potentially highly cost-effective as it uses the infrastructure and

[8] For more details on the training see Appendix 0.

human resources already existing in the health network with no further monetary and organizational costs beyond training and printing material. As outlined in detail in section 7, the unit labor cost of NEP-B is estimated to be at around 62 USD per family attended.[9] The costs for NEP-B are only 15-20% of the costs of home visits. The more intensive version of the program (NEP-I) costs 35% more than the standard version (NEP-B) per family attended.

3 The Evaluation Design

NEP was implemented across Chile. Therefore, our study is based on a representative sample of health clinics located in both urban and rural areas all over the country. The sample was stratified by type of clinic, which included family health centers, general health centers, and small hospitals (this stratification was motivated by the idea that different infrastructure and human resources across types of health centers may play an important role in the delivery of the program).

3.1 Recruitment

To recruit study participants, the research team worked in close collaboration with the Ministry of Health and the health centers that were part of the impact evaluation study to form a list of potential participants in each health center. The outreach to potential households was carried out through the standard recruitment strategies for NEP by facilitators, through a personal interview during the regular health checks, a home visit, or a meeting with other potential participants. In the period from April to June 2011, facilitators constructed a waitlist of about 45-60 families per health center that satisfied both the inclusion and exclusion criteria of eligibility for NEP. The initial list of 45-60 families was drawn to be able to form three potential groups of 10-12 households. Families with higher level of adversities such as suspected domestic violence, severe mental health problems, or child developmental delays requiring clinical attention, were excluded from the group sessions and referred to services with more individualized attention. Once identified, parents of eligible families were invited to participate through home visits or through a direct recommendation made by a health professional. Eligible families were enrolled after an interview with a NEP facilitator, where they were informed about this study intentions and about the randomized process of assignment to groups. Also, they were given the chance to read and sign the informed consent form (or received an assisted reading of the same when they declared difficulties to read, poor reading skills, or illiteracy). Very few parents refused to participate.

3.2 Randomization

After uploading the waitlists on the study online platform, a two-stage randomization process was implemented by the research team. First, we restricted the sample size of the survey data collection to 18 families per health clinic due to resource constraints: a sample of 18 families was randomly drawn from the waitlist to be administered the baseline survey. Second, the 18 families selected to be administered the baseline survey were randomly assigned to three groups: 1/3 was invited to participate in NEP-B, 1/3 was invited to participate in NEP-I, and the remaining 1/3 of families were assigned to the control group. The control group remained on a waitlist until the endline survey was conducted, at which point they became eligible to participate in NEP. Families in the control group receive no NEP benefits, but they continued to receive their usual health care at the health

[9] The estimated cost includes labor for the facilitators for the sessions (for an average of 7 sessions, 2 hours each).

center, which included non-structured information sessions about health and child development with the parents and regular control visits to children. Treatment families were free to accept or not the invitation to participate in NEP. The survey data was collected by an independent survey company with experience collecting public health surveys in Chile. Interviewers knew that the survey was part of the evaluation of the program *Nadie es Perfecto,* but they were blind to treatment assignment.

The final sample includes 162 health clinics stratified by type of health center, 324 facilitators (162 for the basic NEP and 162 for the enhanced NEP-I), and about 18 households per health center (6 NEP-B, 6 NEP-I, 6 control), which resulted in a total sample size of 2,916 caregivers and 3,597 children at baseline.

3.3 Measurements

There are two survey waves used in this study: a baseline survey which occurred before the intervention took place, administered in June-September 2011, and an endline survey administered in July-October 2014, almost three years after the end of the group sessions for the sample of households participating in this study. The 6-to-8-week NEP program occurred in slightly different periods in each participating clinic, and they all occurred between October 2011 (start date for the first NEP group in the study) and April 2012 (end date for the last NEP group in the study). These surveys cover in detail different dimensions of caregiver characteristics and behaviors as well as child developmental outcomes, which we now describe.[10]

3.3.1 Child development measures

We consider three developmental domains potentially affected by the intervention: language, executive function, and socio-emotional development.

Language: At baseline, we measured both receptive and expressive language for children from 0 to 71 months using the Spanish version of the Preschool Language Scale (PLS-4). However, because a large proportion of children at endline were older than 71 months and would have aged out of this test, we complemented the PLS-4 with the "Test de Vocabulario en Imágenes" (TEVI-R), a direct assessment for receptive vocabulary that has been adapted from the Peabody Picture Vocabulary Test and normed for the Chilean context, which was administered at endline to children 36 months of age and older (Echeverría, Herrera, and Segure 2002).[11]

Executive function: These are the cognitive aspects of self-regulation (Blair and Razza 2007), typically encompassing domains of working memory, inhibitory control, attention, and cognitive flexibility. We administered the Dimensional Change Card Sort (DCCS) task (Zelazo 2006), which is a test of cognitive flexibility appropriate for children aged 2½ years and older. In the standard version of the test, children are asked to order a series of cards according to one dimension (for example, the color), and then according to another dimension (for example, the shape). The test requires holding two pieces of information in mind and at the same time inhibiting a dominant tendency when the task is switched. At endline, we also administered a Leiter-R scale to measure the capacity to sustain attention (Roid and Miller 1997).

[10] As we explain below, for our main estimates of program impacts we consider summary indices of these detailed measures constructed from factor models. This enables us to account for measurement error in these measurements, reduce the number of hypotheses we test simultaneously, and estimate parsimonious mediation models.

[11] A subset of children older than 36 months and younger than 71 months of age were administered both the PLS-IV and the TEVI-R. The two measures align well for this subset.

Socio-Emotional Development: We measure positive dimensions of children's socioemotional development (adaptive behaviors) using the Battelle Developmental Inventory Screening Test (BDIST II) Personal-Social Scale (Ringwalt 2008), which includes three sub-domains of socio-emotional development: interaction with adults, interactions with peers, and the self-concept and social role. The first two subscales of BDIST II are available for children up to 71 months (5 years and 11 months), whereas the latter is available for children up to 83 months of age (6 years and 11 months). To measure a wide range of behavioral problems (maladaptive behavior) we administered the Achenbach Child Behavior Checklist (CBCL) (Achenbach and Ruffle 2000), which captures internalizing and externalizing behavioral problems for children aged 1½ years and older. Both the Battelle and CBCL measures behaviors are reported by the primary caregiver.

3.3.2 Parental behaviors

To measure parenting behaviors and home environments we combine self-reported and directly observed variables. In the baseline survey, we administered the Family Care Indicators (FCI) (Hamadani et al. 2010), which measures the frequency of learning and play activities with children, as well as the amount and variety of play and learning materials available at home. At endline, we used again the FCI complemented with additional self-report and observational items from the HOME-Short Form (Bradley and Caldwell 1984), enabling us to expand the Family Care Indicators for children of older age groups. An exploratory factor analysis of these items, which were highly correlated to each other, revealed a single relevant latent factor of cognitive stimulation (labeled as the HOME Index).

In addition, we use two sub-scales of the Parent Behavior Checklist (Fox 1994), where parents were asked to indicate how frequently they engaged in different activities with their child over the past couple of weeks. The first sub-scale measures nurturing practices, associated with positive parental socio-emotional interactions with the child. The second sub-scale measures discipline practices, a mixture of positive and harsh disciplinary practices. An exploratory factor analysis of the nurturing subscale revealed a single relevant latent factor (which we label as the Nurturing Index), and a similar analysis of the discipline subscale revealed two relevant latent factors (labeled as a Negative Discipline Index and a Positive Discipline Index).

3.3.3 Parental beliefs, attitudes, and expectations

At least one-third of the sessions in NEP aim to promote participants' self-care and self-image as parents. This dimension of parental perceptions, related to parental self-efficacy, is grounded in the social cognitive theory (Bandura 1986, Bandura 1995). The premise is that parents cannot be effective if they do not have a strong self-image. In order to measure this construct, we use the Parenting Sense of Competence Scale (PSCS) (Ohan, Leung, and Johnston 2000), a 17-item scale that evaluates parental confidence in their capacity to overcome daily child-rearing tasks. A complementary instrument captures how parents perceive their own behaviors would impact their children's development. To this end, we use a subscale of the Parental Cognitions and Conduct Toward the Infant Scale (PACOTIS) (Boivin et al. 2005), a 5-item Likert scale to assess the perceived parental impact of their behavior on the developing child. We dichotomized the items and constructed a perceived impact indicator by adding all the items.

Social support has been emphasized as an important mediator of change in group-based health interventions (Briscoe and Aboud 2012). To measure perceived social support by parents, we used a short version of the Parental Social Support Scale (PSSS) (Cutrona and Troutman 1986), which

includes subscales for perceived support from the family, from friends, or the community, and from significant others.

To capture parental beliefs about how to raise children, in particular ideas about structure and warmth in child-rearing tasks, we adapt the Ideas About Parenting (IAP) questionnaire (Heming, Cowan, and Cowan 1990). This scale can be used to characterize parenting in terms of three styles: authoritarian, authoritative, and permissive, and we construct raw scores for each of these three sub-scales (Baumrind 1968, Maccoby, Kahn, and Everett 1983).

3.3.4 Maternal mental health and endowments

We collected data on symptoms of depression using the Center for Epidemiologic Studies Depression Scale (CES-D) (Knight et al. 1997), as well as measures of maternal distress with the Parenting Stress Index (PSI) sub-scale Parenting Distress (Abidin 1990). We also administered two scales of the Wechsler Adult Intelligence Scale (WAIS-IV) to caregivers (Vocabulary and Digit Span). This allows us to control for maternal IQ, which is an important predictor of children's cognitive skills. In addition, we also measure the caregiver's personality traits using the Big Five Inventory test (John, Donahue, and Kentle 1991, Goldberg 1993), which assesses extraversion, agreeableness, conscientiousness, openness, and neuroticism.

Finally, we collected socioeconomic data for all the household members including education attainment, age, labor and non-labor income, family composition, employment status, household wealth, access to health and community services, and health shocks.

3.3.5 Construction of summary indices

Rather than analyzing program impacts for individual overlapping measures of child and parental outcomes, our main results rely on the estimation of single indices for the groups of measures mentioned above.

Starting with child development outcomes, we used factor models[12] to construct the following indices of developmental domains:

Vocabulary Index: Estimated using raw data on item responses to the TEVI-R test.

Executive Function Index: Estimated using the raw cognitive flexibility score obtained from item responses to the DCCS test and the raw sustained attention score obtained from item responses to the Leiter-R test.

Socioemotional Index: Estimated using the raw scores of the three sub-scales of the Battelle Socio-personal screening (interaction with adults, interactions with peers, and the self-concept and social role), as well as the raw scores of internalizing and externalizing behaviors as measured by the CBCL scale.

For parental outcomes, we used factor models to create three indices to capture three different

[12] For each domain of child development we estimate factor models assuming normality of the factors and of the errors (we do the same for the different dimensions of parenting described below). Child development measurements can be discrete or continuous. We also estimated a combined Child Development Index using all measures collected at endline. However, this single index was highly correlated with the Socioemotional Index, so we left it out of our results. The methods used to estimate child and parental indexes are described in detail in Appendix 2, and summary indexes with their corresponding measures are presented in Table A5.

constructs:

Parental Behaviors Index: we first estimated a HOME index, a Nurturing Index, a Negative Discipline Index, and a Positive Discipline Index using individual item responses from the FCI scales and PBC scales. We then constructed a composite Parental Behaviors Index using these four indices as measurements.

Parental Beliefs Index: Estimated using raw scores for parental perceived self-efficacy (PSCS), perceived social support from family, friends, and others (PSSS), perceived impact of own behaviors on child development (PACOTIS), as well as the raw scores of three parenting styles (authoritative, authoritarian or permissive) obtained from the IAP scale.

Parental Well-being Index: Estimated using raw scores for parental distress (PSI) and depression (CES-D).

4 Data

4.1 Baseline descriptive and sample balance

The NEP sample at baseline consisted of 2,916 households where the principal caregiver was interviewed, and 3,597 children were assessed with developmental tests. Table 1 describes some key characteristics for our sample, as well as a benchmark with a nationally representative sample of children 0-5 and their caregivers (from the *Encuesta Longitudinal de Primera Infancia*, or ELPI). The first column provides summary statistics from the ELPI, the next three columns are drawn from the NEP sample, one for each treatment arm. We show the mean of each variable and the number of observations in each group. The last two columns of the table display p-values for tests of equality of each treatment arm relative to the control group.

Caregivers are mostly mothers (94.8%), followed by grandmothers (3.6%). Fathers are the main caregiver only in 1.2% of all households. This is consistent with what we see in the administrative records from the program. The NEP sample is on average younger than the national population, with an average age of caregivers of 29 years old, and with most caregivers being between 21 and 30 years of age. In terms of education, 37.5% of caregivers are high school dropouts, and 16.3% have some level of tertiary education. The national sample has a larger fraction of caregivers with secondary and tertiary education.

The intervention targets the most disadvantaged section of the population in Chile. Of the households in the sample, 41.6% belong to the bottom quintile and 66.2% belong to the bottom two quintiles of the per-capita household income distribution in the country (which roughly coincides with the definition of poverty in Chile in 2011). Among participant households, 41.3% are single-mother households (compared to 1/3 in the ELPI), while the remaining are bi-parental nuclear families (consisting of a father, mother, and children, but no other adults at home).

Table 1 also presents the main descriptive statistics at baseline for children participating in the evaluation. Among them, 53.4% are males, with an average age of 27.96 months. About half of the sample of children are below 2 years old at baseline (47.4%), more than half are the first child born, and 1/3 of them are the second born.

We find no significant differences in household, caregiver, or child characteristics (including the children's mean age) across treatment arms at baseline, whether we test for equality of these

characteristics individually or jointly (last line of the table). However, to control for potential imbalances across specific age groups all our treatment effects control for children's age and gender.

In Appendix 1, we show that the baseline sample is also balanced across treatment arms for measures of child development, parenting behaviors and beliefs, and parental mental health. We also show that children's cognitive outcomes, such as language and executive function, are strongly positively correlated with maternal education, and children of more educated caregivers are less likely to exhibit behavioral problems (Fernald et al. 2012, Luoto et al. 2021). These trends are mirrored by measures of parental behaviors and beliefs. Relative to less educated caregivers, those who are more educated provide more cognitive stimulation to their children, are more nurturing, use less harsh disciplinary practices, have a higher perception of self-efficacy, and see child rearing in a more authoritative way, as opposed to authoritarian or permissive parenting styles.

Table 1: Baseline balance, Household and Child Characteristics

	ELPI	NEP				
	(1) Overall	(2) Control	(3) NEP-B	(4) NEP-I	t-test p-value	t-test p-value
Household Characteristics at Baseline	Mean	Mean	Mean	Mean	(1)-(2)	(1)-(3)
Caregiver is the mother (%)	93.3%	94.5%	95.3%	94.7%	0.920	0.994
Caregiver's Education (%)						
Primary	17.0%	20.0%	22.4%	20.1%	0.155	0.908
Secondary Incomplete	16.5%	17.1%	17.0%	15.5%	0.947	0.430
Secondary Complete	40.8%	46.7%	44.3%	47.5%	0.175	0.621
Tertiary	25.6%	16.2%	16.2%	16.8%	0.834	0.792
Single Mother	31.1%	39.3%	41.1%	41.5%	0.266	0.090
Caregiver's Age	29.9	29.1	28.7	29.1	0.834	0.586
	(7.59)	(8.35)	(8.56)	(8.41)		
Household p.c. income (<40%)	59.0%	66.8%	65.9%	67.4%	0.647	0.865
Household p.c. monthly income	232.7	174.6	167.4	173.9	0.240	0.602
(in 2010 US dollars)		(180.15)	(124.20)	(159.81)		
No. Observations		972	972	972		
Child Characteristics at Baseline						
Boys	51.20%	53.4%	52.7%	53.9%	0.686	0.379
Age in months	29.8	28.0	27.9	27.9	0.858	0.460
	(20.26)	(18.63)	(17.93)	(18.41)		
Birth Order						
First		52.3%	55.9%	53.6%	0.193	0.740

Second		31.1%	29.5%	32.2%	0.479	0.591
Third or more		16.5%	14.6%	14.2%	0.374	0.245
No. Observations		1210	1185	1183		

Note: The first column is obtained from the nationally representative ELPI survey (using sampling weights). Columns 2-4 are summary statistics from the NEP sample at baseline (households: N=2916; children: N=3578). The last two columns provide comparisons between the control arm and NEP-B and NEP-I. Household per capita income reported in 2011 US dollars per month. Significance levels: *p<=10%, **p<=5%.

4.2 Attrition

At endline, we were able to interview 2,545 caregivers and 2,895 children, representing a 12.7% (15,8% Control, 10.6% NEP-B, 11.7% NEP-I) and 19.1% (21,7% Control, 16.7% NEP-B, 18.8% NEP-I) of sample attrition across survey waves. Table 2 compares the same characteristics measured at baseline between NEP-B and NEP-I and the control group, but now for the sample of households and children who were interviewed again at endline. All differences in observable characteristics are not statistically different from zero so we find no evidence on selective attrition based on these variables. However, a further examination of outcomes at baseline shows there is slightly more attrition in the control group than either of the treatment groups, particularly NEP-B, but that the differences are small (Appendix 4). We also find that children from the NEP-B arm with higher language scores at baseline were more likely to leave the sample, as well as families from the NEP-I arm with higher incomes.

These sources of potential selective attrition are discussed in Appendix 4 where we present three methods to account for selective attrition. One is the estimation of Lee Bounds (Lee 2009). The second is the use of Inverse Probability Weighting (IPW) (Robins, Rotnitzky, and Zhao 1994) to reweight our data in such a way that a larger weight is given to participants who are underrepresented in the endline sample as a result of attrition. The third method censors the outcomes of interest at different values and examines how estimates change (Angrist, Bettinger, and Kremer 2006). Our results are robust to these corrections for attrition.

Table 2: Balance of Household and Child Characteristics at Follow-up

	(1) Control	(2) NEP-B	(3) NEP-I	t-test p-value	t-test p-value
Household Characteristics at Baseline in the Follow-up sample	Mean	Mean	Mean	(1)-(2)	(1)-(3)
Caregiver is the mother (%)	94.7%	94.9%	94.5%	0.955	0.844
Caregiver's Education (%)					
Primary	20.0%	22.6%	20.0%	0.168	0.969
Secondary Incomplete	17.1%	16.8%	15.4%	0.834	0.369
Secondary Complete	47.2%	44.8%	48.5%	0.287	0.638
Tertiary	15.6%	15.9%	16.1%	0.890	0.815

Single Mother	40.1%	41.7%	43.2%	0.408	0.152
Caregiver's Age	29.0	28.7	29.1	0.676	0.686
	(8.27)	(8.72)	(8.40)		
Household p.c. monthly income	173.4	164.5	176.0	0.188	0.730
(in 2010 US dollars)	(189.07)	(114.55)	(163.89)		
No. Observations	818	869	858		
Child Characteristics at Baseline in the Follow-up sample					
Boys	52.1%	52.5%	54.1%	0.686	0.379
Age in months	28.2	28.0	27.3	0.858	0.460
	(18.33)	(17.93)	(18.11)		
Birth Order					
First	52.3%	55.9%	53.6%	0.193	0.740
Second	31.1%	29.5%	32.2%	0.479	0.591
Third or more	16.5%	14.6%	14.2%	0.374	0.245
No. Observations	947	987	961		

Note: Columns 1-3 are summary statistics of baseline measures from the NEP sample at follow-up (households: N=2545; children: N=2895). The last two columns evaluate selective attrition based on each variable by providing comparisons between the control arm and NEP-B and NEP-I. Household per capita income reported in 2011 US dollars per month. Significance levels: *p<=10%, **p<=5%.

5 Methods

Our primary parameters of interest is given by intent to treat estimates (ITT) of the impacts of NEP-B and NEP-I on child and parental outcomes. We estimate the following model for outcome *k*:

$$Y_{i,c,end}^{k} = \alpha_0 + \alpha_1 Z_{i,c}^{B} + \alpha_2 Z_{i,c}^{I} + \Gamma_c + X_{i,c}\alpha_3 + \alpha_4 Y_{i,c,bas}^{k} + \varepsilon_{i,c} \qquad (1)$$

In equation (1), $Y_{i,c,end}$ is an outcome of interest measured at endline, $Z_{i,c}^{B}$ and $Z_{i,c}^{I}$ are indicators for being offered NEP-B and NEP-I respectively, Γ_c represents health center fixed effects that capture unobservable differences in program quality, X_t is a set of control variables including children's attributes such as sex and age in the base specification (or household characteristics such as family structure, household's per capita income, caregiver's education), and $Y_{i,bas}$ is the outcome measured at baseline. Coefficients α_1 and α_2 are the ITT parameters of interest.

Since participation in NEP is voluntary, compliance with treatment assignment can be imperfect. Non-compliance can take the form of individuals invited to NEP not attending the sessions, or individuals assigned to the control group that make it to attend the sessions anyway. In this case,

we complement the ITT analysis with instrumental variable (IV) estimation of program impacts. The first and second stage equations for this estimator are:

$$D_{ic}^{B} = \alpha_0 + \alpha_1 Z_{ic}^{B} + \nu_{ic}^{B} \quad (2)$$

$$D_{ic}^{I} = \gamma_0 + \gamma_1 Z_{ic}^{I} + \nu_{ic}^{I} \quad (3)$$

$$Y_{i,c,end}^{k} = \beta_0^k + \beta_1^{B,k} D_{i,c}^{B} + \beta_2^{I,k} D_{i,c}^{I} + \Gamma_c^k + X_{i,c}\beta_3^k + \beta_4^k Y_{i,c,bas}^{k} + v_{ic}^{k} \quad (4)$$

where the random assignment indicators Z_{ic}^{B} and Z_{ic}^{I} are instruments for participation in each treatment arm, D_{ic}^{B} and D_{ic}^{I}. The randomization is a good IV to control for selection bias into participation in the program as it satisfies both the exclusion restriction and is relevant to predict participation (as we demonstrate below in Table 5). Since there is imperfect compliance, the parameters β_1 and β_2 identify the Local Average Treatment Effects (LATE), which is the average impact of the program for the subgroup of people who comply with their treatment assignment.

As we document below, the proportion of parents taking up the program in our sample (conditional on getting an offer) is below 30%, which means that the LATE parameter could be quite different from the average impact of the program in the population. It may correspond to the average treatment on the treated, since non-compliance among households who were not offered the program was quite small (5%). Since our data comes from the initial stages of the program, one could worry that the composition of the population being served changes over time, especially if the program became more popular leading to higher take up rates. We do not expect this to be the case, as the take-up rate of the program from administrative data has been quite stable over time.

The outcomes we consider are measures of child development, as well as measures of parental beliefs, behavior, and well-being. Since the program is expected to work by changing parental behaviors, it is natural to also present a standard mediation analysis to examine to what extent changes in parental outcomes can account for the impacts of the program on child outcomes. It is necessarily an exploratory analysis given that we cannot interpret the mediation model as a causal model, because there exist potentially important mediators that we do not observe. Furthermore, since mediators and outcomes are measured at the same time, we cannot rule out that there is possible reverse causality from the outcome to the mediators.

For simplicity, we conduct the mediation analysis only on the ITT parameters. In addition to estimating equation (1) for each child outcome *k*, $Y_{i,c,end}^{child,k}$, and each parental outcome *l*, $Y_{i,c,end}^{parent,l}$, we also estimate the following set of equations:

$$Y_{i,c,end}^{child,k} = \pi_0 + \pi_1 Z_{i,c}^{B} + \pi_2 Z_{i,c}^{I} + \Gamma_c + X_{i,c}\pi_3 + \pi_4 Y_{i,c,bas}^{child,k} + \sum_l \pi_5^l Y_{i,c,end}^{parent,l} + w_{i,c} \quad (5)$$

$$Y_{i,c,end}^{parent,l} = \beta_0^l + \beta_1^{B,l} D_{i,c}^{B} + \beta_2^{I,l} D_{i,c}^{I} + \Gamma_c^l + X_{i,c}\beta_3^l + \beta_4^l Y_{i,c,bas}^{parent,l} + v_{ic}^{l} \quad (6)$$

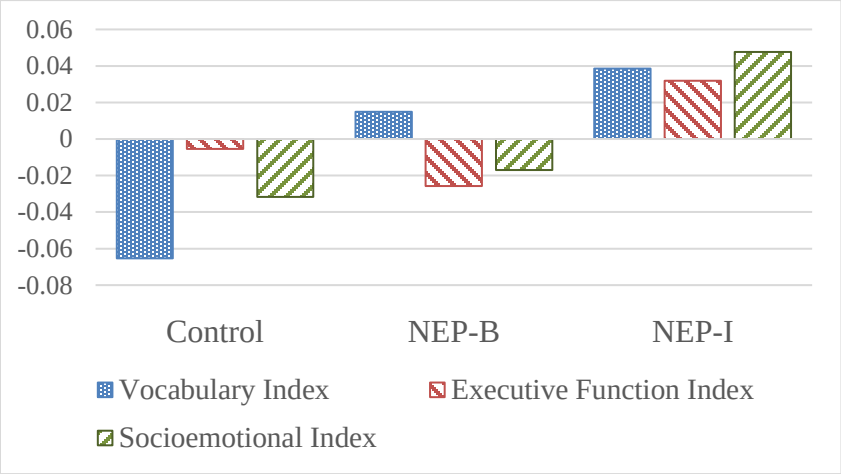
Using this model, we can examine how π_1 and π_2 change in equation (5) relative to α_1 and α_2 in equation (1), when we include parental outcomes as regressors. In addition, under very strong assumptions (orthogonality between $Y_{i,c,end}^{parent,l}$ and $w_{i,c}$ given all the other regressors), we can

interpret $\beta_1^{B,l} * \pi_5^l$ and $\beta_2^{I,l} * \pi_5^l$ as the component of the overall intent to treat effect (α_1 and α_2 in equation (1)) coming through $Y_{i,c,end}^{parent,l}$, which is often denominated as the intervention's "indirect effect". Finally, we can also test if the mediation parameters are invariant to treatment assignment by including interaction terms between the mediation variables and treatment assignment in equation (6).

6 Results

This section presents the estimated impacts of NEP-B and NEP-I on child outcomes and parenting behaviors and beliefs. Because the endline survey was collected from households in our sample between 30 to 36 months after the end of the interventions, our results can be interpreted as medium-term effects of the program.

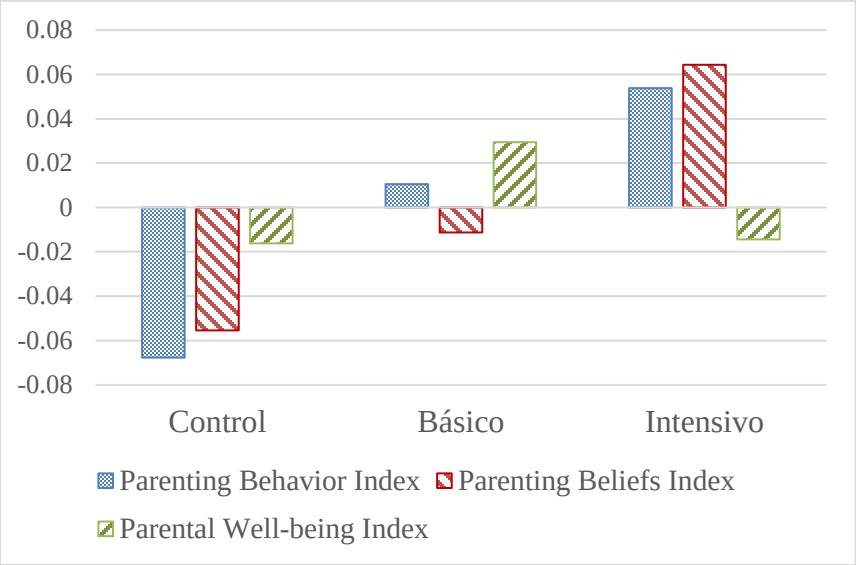
Figure 1: Child Development Outcomes at Follow-up by Treatment Arm



Note: Mean child development indices by treatment assignment status at follow-up.
Latent indices are age-standardized and constructed as indicated in section 3.3.5.

Figure 1 displays average values of our three child development indices (Vocabulary, Executive Function, and Socioemotional) measured at endline, by intervention arm. Each index is standardized to have mean zero and standard deviation one in the sample. The figure is not controlling for the study design effect. Nonetheless, it is suggestive of NEP's positive impacts on child development, at least for children's vocabulary and socioemotional development. Children in the control group had the lowest values of the outcomes, those whose parents were offered the chance to enroll in a NEP-I group had the highest outcomes, and those in the NEP-B groups were in the middle. The exception is the Executive Function Index, which is slightly higher in the control group than in the NEP-B group. The largest difference observed in the figure is between children's vocabulary scores in the NEP-I and control groups, which is of about 0.1 SD.

Figure 2: Parental Outcomes at Follow-up by Treatment Arm



Note: Mean parental indices by treatment assignment status at follow-up. Latent indices are standardized and constructed as indicated in section 3.3.5.

Figure 2 presents the analogous picture for our three parental indices: parenting behaviors, parenting beliefs, and parental well-being. The pattern is similar to the one in Figure 1, where the largest values of the indices of parenting behaviors and beliefs are for those randomly assigned to being offered NEP-I, followed by those randomly assigned to being offered NEP-B, and then those in the control group. Again, this indicates that NEP induced important changes in parenting behaviors and beliefs. Differences in parental well-being are small across the three groups.

In the remainder of this section, we present estimates of program impacts for the main specification, where we only control for child age and gender, and for health center fixed effects, as well as for an extended specification that includes household characteristics, maternal cognition and personality traits, and children's baseline outcomes. Our results are robust to the inclusion of all these controls, as well as to controlling for interviewer fixed effects and clustering standard errors at the household level (Appendix 3). We supplement standard inference procedures with multiple hypothesis testing (Romano and Wolf 2005), Click or tap here to enter text.where child and parental outcomes are tested simultaneously.

6.1 Intention-to-Treat

Table 3 shows our main ITT estimates of the impacts of NEP-B and NEP-I on age-standardized indices of child development, based on equation (1). The main specification (Column 1) only controls for children's age and gender, and for health center fixed effects. Column 2 adds caregiver's and household's characteristics, and column (3) adds an index of baseline child development outcomes estimated with factor models that include all children's cognitive, language and socioemotional measures collected at baseline. In the main specification, our results show that the offer of participation in NEP-B led to improvements in child vocabulary, but which are not robust across specifications and to multiple hypothesis testing. In addition, we do not find significant impacts of offering NEP-B on children's executive function or socio-emotional skills.

In contrast, offering households the chance of participating in NEP-I led to improvements in the Vocabulary Index by 0.099 SD and the Socioemotional Index by 0.094 SD, relatively to the control group. These improvements are statistically significant at the 5% level and robust to multiple hypothesis testing of correlated outcomes also at the 5% level. We found no significant impacts in the Executive Function Index in any treatment arm. The coefficients are almost unchanged when more controls are added in columns 2 and 3.[13] These impacts are similar to the suggested magnitudes shown by Figure 1.[14]

Table 3: ITT estimates of child development outcomes

		(1)		(2)		(3)	
Child Outcome	Obs.	NEP B	NEP I	NEP B	NEP I	NEP B	NEP I
Vocabulary Index	2895	0.070	0.099**††	0.076*	0.103**††	0.075*	0.103**††
		(0.044)	(0.045)	(0.044)	(0.045)	(0.044)	(0.045)
Exec. Function Index	2895	-0.021	0.042	-0.024	0.039	-0.025	0.040
		(0.044)	(0.044)	(0.044)	(0.044)	(0.044)	(0.044)
Socioemotional Index	2492	0.032	0.094**††	0.040	0.094**††	0.041	0.096**††
		(0.047)	(0.047)	(0.047)	(0.047)	(0.046)	(0.046)
*Child age and gender, health center FE		Y	Y	Y	Y	Y	Y
*Caregiver/household characteristics		N	N	Y	Y	Y	Y
*Baseline outcomes		N	N	N	N	Y	Y

Note: Each line reports estimates from a separate regression. Dependent variables are indices for child developmental outcomes measured at follow-up estimated with Generalized Simultaneous Equation Modelling (GSEM) using raw data, except for the Vocabulary index which was estimated with IRT models. Effect sizes use internal age-standardization to the full sample (mean 0, SD=1). All regressions control for children's age and sex, and for health center's fixed effects. Regressions in columns (2) control also for caregiver characteristics (educational attainment, IQ and personality traits) as well as household characteristics (household income and size). Regressions in column (3), in addition, control for an index of baseline child development outcomes. Significance levels: *p<=10%, **p<=5% for individual hypotheses tests; ††p<=5% †p<=10% after accounting for multiple hypotheses tests.

Table 4 presents the ITT estimates for caregiver indices of parenting behaviors, beliefs, and psychological well-being. The main specification (Column 1) controls for children's age and gender, as well as health center fixed effects, and column 2 adds caregiver and household characteristics. In our main specification, the offer to participate in NEP-I significantly improved caregivers' parenting behaviors by 0.108 SD, reflecting improvements in the quality of the home environment, as well as nurturing, negative and positive discipline practices. We also found that

[13] The results are robust to further sensitivity analyses including interviewer fixed effects and clustering SEs at the household level, as there are a few households for which we observe more than one child (Appendix 3).

[14] Despite seemingly large differences in impacts across treatment arms, we can never reject the hypothesis of equal impacts of NEP-B and NEP-I for any outcome and specification. Estimated effects for each of the components used to construct developmental indexes and behavior and beliefs indexes are also shown for reference in Table A7 and Table A8 in Appendix 3.

NEP-I significantly improved caregiver's parenting beliefs by 0.111 SD, which is the result of improvements in parental perceived self-efficacy, perceived impact of own behavior on child development (PACOTIS) and perceived social support from friends (Table A8). We find no significant impacts of NEP in psychological well-being including measures of caregiver mental health (CES-D) or stress (PSI), in line with international evidence (Waldrop et al. 2021, Jeong, Pitchik, and Fink 2021). NEP did not include a specific component aimed at improving caregivers' mental health, nor were NEP facilitators trained to tackle mental health problems. A more targeted program may be needed to shift these outcomes.

All our results are robust to the inclusion of controls such as caregivers' education, IQ, and personality traits, as well as household income and composition (Column 2). Moreover, these results remain invariant to the inclusion of interviewer fixed effects or clustering at the household level (Appendix 3).

Impacts of NEP-B on parenting behaviors, parenting beliefs, and parental well-being are not statistically different from zero. Nevertheless, as in the case of child outcomes, we can never reject the hypothesis of equal impacts of NEP-B and NEP-I for any outcome in any specification.

Table 4: ITT estimated parameters parental practices and parental beliefs

		(1)		(2)	
Parental Outcome	Obs.	NEP B	NEP I	NEP B	NEP I
Behavioral Index	2545	0.063	0.108**††	0.067	0.096**††
		(0.046)	(0.046)	(0.044)	(0.044)
Beliefs Index	2545	0.037	0.111**††	0.027	0.093**††
		(0.047)	(0.047)	(0.042)	(0.042)
Well-being Index	2545	0.047	0.012	0.042	0.015
		(0.046)	(0.047)	(0.042)	(0.042)
*Child age and gender, health center FE		Y	Y	Y	Y
*Caregiver/household characteristics		N	N	Y	Y

Note: Each line reports estimates from a separate regression. Dependent variables are caregiver outcomes measured at follow-up. The Behavioral index was estimated with GSEM combining estimates of the HOME score, as well as nurturing, negative discipline and positive discipline scores, also estimated with GSEM. The Beliefs Index was estimated with GSEM models of parental self-efficacy, perceived impact of own behavior on child outcomes, social support and parenting styles. The Well-being index was also estimated with GSEM using raw scores for parental stress and depression. Effect sizes use internal age-standardization to the full sample (mean 0, SD=1). All regressions control for child sex and age, and for health center's fixed effects. Regressions in columns (2) control also for caregiver characteristics (educational attainment, IQ and personality traits) as well as household characteristics (household income and size). Significance levels: *p<=10%, **p<=5% for individual hypotheses tests; ††p<=5% †p<=10% after accounting for multiple hypotheses tests.

Our estimates from tables 3 and 4 combined indicate that the program was effective in improving child vocabulary and socioemotional outcomes as well as parenting behaviors and beliefs in the medium term. Looking at the components of parent and child indices, our impacts on indices of child outcomes are driven by improvements in receptive language skills and in socioemotional skills such as children's concept of themselves and interactions with adults (Table A7). These changes are mirrored by impacts of similar magnitude in parenting behaviors related to the quality of the home environment and to nurturing and positive discipline practices, as well as by parental

beliefs such as parental perceived self-efficacy, perceived impact of own behaviors on child outcomes, and perceived social support (Table A8).

6.2 Instrumental Variables

We now discuss IV estimates of the impact of participating in NEP-B and NEP-I on child and caregiver outcomes. There are different ways of defining participation in the program, since parents may attend more than one but less than the full number of programmed sessions. Defining participation in the program as attending at least one session of the program offered, administrative program records show that the overall participation rate in NEP-B was 24.9% and in NEP-I was 30.8% among eligible individuals.

There was also imperfect compliance in the control group. The original plan was to start offering the program to the control group one year after the start of the study when the endline survey was originally planned. However, given the delay of the endline (it was administered 30-36 months after the end of the interventions, when the original plan was to have it 12 months after the end of the interventions), a small part of the control group eventually received treatment. We find that 4.8% of caregivers assigned to the control group were able to access the program and attend at least one session of NEP-B, while 5.0% of caregivers in the control group attended at least one session of NEP-I. While our main results focus on this binary measure of participation, we also consider alternative measures of participation such as the number of sessions attended.

6.2.1 Program Participation

Table 5 (columns 1 and 3) describes estimates of the regressions of program participation on indicators for randomly assigned treatment arm when no covariates are included. Column 1 corresponds to equation (2), and Column 3 corresponds to equation (3). The impacts of being offered a slot in NEP-B and NEP-I on participation in these programs were 20.1% and 26.0%, respectively.

Columns 2 and 4 of Table 5 add controls to the estimation of equations (2) and (3) and these estimates are hardly affected. Table A9 in Appendix 3 shows the coefficients on these additional predictors of control participation. Caregivers with a child between 25 and 36 months at baseline are 4.2% more likely to attend sessions in NEP-I. We do not observe a significant association between household income and participation in NEP-B, but households belonging to the second income quantile are 3.9% more likely to attend NEP-I than those at the bottom of the income distribution. The likelihood of participation is higher among more educated caregivers in NEP-B, but it is not relevant to explain participation in NEP-I. Single mothers are 3.4% less likely to attend NEP-B and 3.2% less likely to attend NEP-I, and caregivers who were employed at baseline are 3.0% less likely to participate in NEP-B and 2.5% less likely to participate in NEP-I. Taking the last two indicators together, the data suggests that there were important time constraints on participation among working caregivers with little childcare support.

Finally, the average number of sessions attended by participants was 5.68 sessions in NEP-B and 7.89 sessions in NEP-I. Therefore, the estimated impact on participants reported in Tables 6 and 7 in the next sub-section can be interpreted as the average impact of these number of sessions in each treatment arm.

Table 5: Program take-up

	Participation NEP-B		Participation NEP-I	
	(1)	(2)	(3)	(4)
NEP-B	0.201***	0.200***		
	(0.012)	(0.012)		
NEP-I			0.260***	0.262***
			(0.013)	(0.013)
Controls	N	Y	N	Y
N	2,545	2,530	2,545	2,530

Note: Columns 1 and 3 control only for health center fixed effects. Columns 2 and 4 add households' socio-demographic characteristics as well as caregiver's labor status at baseline. Significance levels: *p<=10%, **p<=5%, ***p<=1%

6.2.2 Impact on program participants

Table 6 shows the IV estimates of the impact of participating in NEP-B and NEP-I on child outcomes, corresponding to equation (4). As in Table 3, participation in NEP-B did not lead to robust improvements in child outcomes under any specification. In contrast, in our main specification (Column 1), participation in NEP-I led to an increase in the Vocabulary Index by 0.43 SD and the Socio-emotional Index by 0.38 SD, results that are statistically significant at the 5% level and robust to multiple hypotheses testing. These impacts are remarkably large given the duration and intensity of the program but need to be taken with caution as they are only applicable to the sub-sample of program compliers and cannot be extrapolated to the whole sample.

There were no statistically significant impacts on the Executive Function Index. These results consistently mirror the ITT analysis after adjusting for participation. When we add more controls (Columns 2 and 3), the IV coefficients on child outcomes remain fairly stable, and standard errors hardly change either. The same happens when we include interviewer fixed effects and cluster standard errors at the household level (Table A10 in Appendix 3).

Our conclusions are robust to multiple hypothesis testing, and to using the number of sessions attended in each program as the measure of participation instead of a binary indicator for attending at least one session. In Table A11 in Appendix 3, we document that participation in an average NEP-I session improved the Vocabulary Index by 0.056 SD per session attended and the Socio-emotional Index by 0.048 SD per session attended, results that are robust to the inclusion of more controls. We also found that participation in an average NEP-B session does not lead to improvements in any child outcome.

Table 6: Effect of participation in NEP on child outcomes

		(1)		(2)		(3)	
Child Outcome	Obs.	NEP B	NEP I	NEP B	NEP I	NEP B	NEP I
Vocabulary Index	2895	0.386	0.425**††	0.417*	0.445**††	0.415*	0.447**††
		(0.243)	(0.189)	(0.242)	(0.189)	(0.241)	(0.189)
Exec. Function Index	2895	-0.107	0.149	-0.125	0.133	-0.130	0.137
		(0.240)	(0.187)	(0.238)	(0.187)	(0.237)	(0.186)
Socioemotional Index	2492	0.181	0.381*††	0.226	0.385*††	0.229	0.394**††
		(0.257)	(0.201)	(0.251)	(0.197)	(0.250)	(0.196)
*Child age and gender, health center FE		Y	Y	Y	Y	Y	Y
*Caregiver/household characteristics		N	N	Y	Y	Y	Y
*Baseline outcomes		N	N	N	N	Y	Y

Note: Each line reports estimates from separate 2SLS regressions using randomization status as instrumental variables for attending at least one session. Dependent variables are indices of child development outcomes measured at follow-up. Effect sizes use internal age-standardization to the full sample (mean 0, SD=1). All regressions control for child sex and age, and for health center's fixed effects. Regressions in columns (2) control also for caregiver characteristics (educational attainment, IQ, and personality traits) as well as household characteristics (household income and size). Regressions in column (3), in addition, control for an index of baseline child development outcomes. Significance levels: *p<=10%, **p<=5% for individual hypotheses tests; ††p<=5% †p<=10% after accounting for multiple hypotheses tests.

As in the ITT analysis, participating in the program also led to sustained and remarkably large changes in parenting behaviors and beliefs. Participation in NEP-I significantly improved our index of parenting behaviors by 0.46 SD and our index of parenting beliefs by 0.46 SD in our main specification (Column 1). We can reject that these estimates are equal to zero, even with multiple hypothesis testing. As in the case of child outcomes, these estimates remain fairly unchanged when we add other controls such as caregiver's and household's characteristics (Column 2), and interviewer's fixed effects (Table A10). Participation in NEP-B did not lead to statistically significant improvements in parenting behaviors and beliefs, or parental well-being.

Finally, using the number of sessions attended as the measure we find that of participation in NEP-I leads to an improvement in parenting behaviors of 0.060 SD per session attended and in parenting beliefs of 0.059 SD per session attended (Table A12).

Table 7: Effect of participation in NEP on parental outcomes

		(1)		(2)	
Parental Outcome	Obs.	NEP B	NEP I	NEP B	NEP I
Behavioral Index	2545	0.353	0.460**††	0.373	0.416**††
		(0.258)	(0.200)	(0.245)	(0.191)
Beliefs Index	2545	0.211	0.457**††	0.156	0.379**††
		(0.263)	(0.204)	(0.229)	(0.179)

Well-being Index	2545	0.259	0.080	0.229	0.085
		(0.263)	(0.205)	(0.234)	(0.183)
*Child age and gender		Y	Y	Y	Y
*Caregiver/household characteristics		N	N	Y	Y

Note: Each line reports estimates from separate 2SLS regressions using randomization status as instrumental variables for attending at least one session. Dependent variables are indices of parental behaviors, beliefs and psychological well-being measured at follow-up and estimated with GSEM methods. Effect sizes use internal age-standardization to the full sample (mean 0, SD=1). All regressions control for child sex and age, and for health center's fixed effects. Regressions in columns (2) control also for caregiver characteristics (educational attainment, IQ, and personality traits) as well as household characteristics (household income and size). Significance levels: *p<=10%, **p<=5% for individual hypotheses tests; ††p<=5% †p<=10% after accounting for multiple hypotheses tests.

Our IV estimates from tables 6 and 7 are consistent with our ITT results but they also suggest that, after adjusting for participation, the magnitude of impacts in child and parental outcomes is very large for the sub-sample of program compliers, which is notable for a program of such low cost and low intensity. In line with our ITT results, IV impacts on indices of child development are driven by child language and socio-emotional measures (Table A13), and changes in indices of parental outcomes are driven by measures of the quality of the home environment, parental nurturing, and positive discipline practices, parental perceived self-efficacy, and parental perceived impacts of own behaviors on their children (Table A14). The fact that both ITT and IV impacts on child outcomes occur simultaneously to changes in many parental measures strengthens our confidence that this program leads to improved home environments.

6.2.3 Understanding program take-up

Our results show that, while participation in the program was rather low, the gains among those who participated were quite large. Moreover, simple comparisons between OLS and IV estimates for participation across key child and parental outcomes (Table A15 in Appendix 3) show that IV estimates are 3 to 4 times larger than OLS estimates. This is due to a combination of the take-up rate being small, and the selection of households into the program. Therefore, it is important to understand who self-selects into the program.

We presented above a description of who takes up the offer of the program based on observable characteristics (Table 5 and Table A9). There was no strong relationship between household income[15] and program participation, and some other variables predict participation in NEP-B but not NEP-I, or vice versa. Exceptions are single motherhood and maternal employment, which seem to be correlated with lower participation in either of the two NEP modalities, perhaps suggesting that time constraints may represent salient barriers to participation into the program. Overall, although there is some selection on observables, it is not large and does not point to a specific group of families for whom NEP impacts are likely to be very large.

We then turn to examine selection on unobservables. Following (Black et al. 2020), who propose simple models to test for selection on unobservables, we estimate the following equations:

[15] This pattern is observed within our sample, which is representative of the target population, although not necessarily representative of the entire population in Chile.|

$$E\left[Y_{i,c,end}\middle|X_{i,c},\Gamma_c^I,Y_{i,c,bas},D_{i,c}^I=1\right]=\varphi_0^I+\varphi_1^I Z_{ic}^B+\varphi_2^I Z_{ic}^I+\Gamma_c^I+X_{i,c}\varphi_3^I+\varphi_4^I Y_{i,c,bas} \quad (7)$$

$$E\left[Y_{i,c,end}\middle|X_{i,c},\Gamma_c^I,Y_{i,c,bas},D_{i,c}^B=1\right]=\varphi_0^B+\varphi_1^B Z_{ic}^B+\varphi_2^B Z_{ic}^I+\Gamma_c^B+X_{i,c}\varphi_3^B+\varphi_4^B Y_{i,c,bas} \quad (8)$$

$$E\left[Y_{i,c,end}\middle|X_{i,c},\Gamma_c^I,Y_{i,c,bas},D_{i,c}^I=D_{i,c}^B=0\right]=\varphi_0^C+\varphi_1^C Z_{ic}^B+\varphi_2^C Z_{ic}^I+\Gamma_c^C+X_{i,c}\varphi_3^C+\varphi_4^C Y_{i,c,bas} \quad (9)$$

Equations (7), (8) and (9) are essentially regressions of the outcome of interest on the random assignment indicators (Z_{ic}^B and Z_{ic}^I) and our set of controls ($X_{i,c},\Gamma_c,Y_{i,c,bas}$), using the sub-samples of participants into either program, or the subsample of nonparticipants, respectively. Black et al. (2020) show that, if we can reject the null hypothesis that any of the coefficients on Z_{ic}^B and Z_{ic}^I in any of the equations is equal to zero, then there must be selection on unobservables.

Results for child vocabulary and socioemotional outcomes are shown in Table 8 (see Table A16 in Appendix 3 for results on parental outcomes). However, in our specific case this test is likely to be most reliable only for equation (9) above (the subsample of non-participants), not only because it has by far the largest sample, but also because most of the non-compliance in our setting is linked to lack of take-up of the program by those who are eligible, rather than take-up of the program by ineligible households which is very uncommon (without the latter one cannot estimate equations (7) or (8) since there is no variation in *Z* given *D*). For completeness, we present the entire set of results, but we focus the discussion on the results from equation (9).

We find suggestive evidence of selection based on unobservables into the intensive program among non-participants (column 3 in Table 8 and Table A16 in Appendix 3). Furthermore, this selection appears to be negative: among those who do not participate in the program, families who were invited to NEP-I show better socioemotional outcomes of children than those in the control group. Overall, the estimates are too imprecise to make strong statements about selection on unobservables, but it is possible that returns are especially large for those with the lowest outcomes and that the magnitude of our estimates of program impacts are large in part because they concern a specific population with potentially low levels of unobservable determinants of child outcomes.

Table 8: Selection on Unobservables

	(1) Participants NEP-B	(2) Participants NEP-I	(3) Non-Participants
Panel A: Vocabulary Index			
NEP-B	0.036	0.096	0.031
	(0.176)	(0.248)	(0.052)
NEP-I	-0.021	0.095	0.081
	(0.241)	(0.191)	(0.053)
p-value joint test NEP-B=NEP-I=0	0.957	0.878	0.313
Observations	343	399	2153
Panel B: Socioemotional Index			

NEP-B	-0.063	-0.058	0.018
	(0.196)	(0.261)	(0.055)
NEP-I	-0.430	0.016	0.114**
	(0.267)	(0.198)	(0.056)
p-value joint test NEP-B=NEP-I=0	0.240	0.939	0.106
Observations	292	350	1850

Note: For each panel (index of interest), each column represents a separate OLS regression. Column 1 uses the sample of participants into the NEP-B program only. Column 2 uses the sample of participants into the NEP-I only. Column 3 uses the sample of non-participants only. Dependent variables are indices of key child and parental outcomes measured at follow-up. Effect sizes use internal age-standardization to the full sample (mean 0, SD=1). All regressions control for child sex and age, and for health center's fixed effects. Significance levels: *p<=10%, **p<=5% testing individual hypotheses.

6.3 Heterogeneous Treatment Effects

We also examined whether the impacts of NEP on child outcomes were heterogeneous along caregiver characteristics such as caregiver education, mental health, cognition, and personality traits, or along child outcomes at baseline, child age, and child gender (we report all these results in Tables A22 to A32 in the Appendix 5). This analysis is largely exploratory. Overall, we found evidence of larger impacts of NEP-I in the Socioemotional Index among lower-educated caregivers (Table A22), and larger effects in the Vocabulary Index among caregivers with lower cognitive abilities (Table A23).

While there is suggestive evidence that the impacts are slightly larger for more disadvantaged families, we do not find support for substantial heterogeneity across other child or parental outcomes. For example, we cannot shed light on the question of when it is better to intervene as we can never reject the null hypothesis that treatment effects on child outcomes are statistically the same across younger and older children in our sample (Table A26).[16] We did not find strong evidence of heterogeneity by facilitator background either (Table A34). Finally, our results show that child socioemotional outcomes are higher in general health centers relative to more specialized types of centers such as family health centers and small hospitals, but we find no evidence of heterogeneous impacts by type of health center for vocabulary and executive functions outcomes (Table A35).

6.4 Mediation Analysis

Our results above show that participation in NEP results in sustained improvements in child vocabulary and socio-emotional development, as well as in indices of parenting behaviors and beliefs. Given the short duration of the program, it is reasonable to think that any program impacts observed 3 years after program completion operate primarily by changing parenting behaviors and beliefs in the long run.

[16] These results are presented in Table A26 in Appendix 5 and are robust to the age cutoff used for the analysis (mean or median age at baseline, 24 or 36 months at baseline as potential cutoffs). If anything, we find that treatment effects in vocabulary are larger for older children, while effects in socio-emotional development are larger among younger children, but these signals are not strong enough to draw conclusions.

In this section, we estimate a standard mediation analysis model to examine to what extent the impacts of NEP on the vocabulary and socioemotional indices can be explained by the impacts of NEP on our indices of parenting behaviors and beliefs.[17] The assumptions under which one can decompose treatment effects estimates into different components are however strong, as discussed above. This means, as usual in this type of analysis, that the results in this section can only be interpreted as suggestive evidence of the importance of these mediators.

Recall our mediation model outlined in the system of equations (5) and (6). The goal is to separately identify the pathways through which NEP could have affected child outcomes: a) indirect effect through changes in the level of parenting behaviors and beliefs (captured by the $\beta_1^{B,l} * \pi_5^l$ and $\beta_2^{I,l} * \pi_5^l$ terms), and b) the direct intervention effect through changes in unobserved inputs (π_1 and π_2). We simplify the analysis by not allowing the productivity of baseline outcomes and household characteristics to change with treatment. In the case of NEP-B, the program was unable to significantly shift any intermediate indicators, so here we focus on NEP-I.

We estimate the model in steps using a Monte Carlo simulation approach following (Campos et al. 2017). First, the coefficients $\beta_1^{B,l}$ and $\beta_2^{I,l}$ are obtained by regressing each mediator on treatment assignment. Second, we obtain estimates of π_5^l from a regression of each child index on treatment status (as in the ITT equation, controlling for child, caregiver, and household characteristics and health center fixed effects) and on each mediator. We could also include interaction terms between mediation variables and treatment assignment to test for the possibility that productivity changes with treatment assignment, but as we discuss above, we ignore this in our main calculations.

Next, we compute the 95% Monte Carlo confidence intervals for the indirect effect of each mediator based on a very large number of repetitions. A confidence interval that does not include zero indicates a significant indirect effect of that particular mediating variable on child outcomes. Finally, for the total indirect effect, we include all the relevant mediators defined as those with statistically significant paths in the model.

Table 9 describes our main results for the Vocabulary Index. Column 1 reports the ITT coefficients of the impact of the program on the outcome (from table 3). Columns (2)-(4) add one significant mediator at the time to the model; column (4) adds all intermediate outcomes that are significantly shifted by the NEP-I. A parental index is a mediator if it statistically explains the main outcome and the confidence intervals for the indirect effect at the bottom of the table never include the zero. In this case, only the parenting behavioral index is a mediator. The total impact estimate is 0.099 SD, which declines to 0.092 SD when we add the significant mediators (column 4). This means that the indirect intervention effect on child vocabulary that is explained by changes in parenting behaviors at most explain about 8% of the effect of the total effect of NEP-I on child vocabulary. When interaction terms between mediators and treatment assignment are added, these are not statistically different from zero. Mediation results using individual parental scores shown in Table A33 in Appendix 6 show that only home environment and nurturing behaviors are relevant mediators for impacts on child vocabulary.

Table 10 presents the results of the mediation analysis for the Socio-emotional Index. In this case, both the parenting behavioral and beliefs indices are relevant mediators and individually each of

[17] In the main analysis, our vector of potential mediators only considers our overall indexes for parental behaviors and parental beliefs. In Appendix 6 we present the analogous mediation results if we included individual measures of parental behaviors and beliefs.

them explains 36.6% and 31.4%, respectively, of the total intervention effects on socioemotional development. When added together, the overall impact of NEP-I decreases from 0.094 SD to 0.050 SD and is no longer statistically significant. That is, the combined effect of parenting behaviors and beliefs explains up to 47.9% of the treatment effects. As with the mediation model for child vocabulary, the estimated interaction terms between mediators and treatment assignment are not statistically different from zero. Mediation results using individual parental scores shown in Table A34 show many relevant mediators including the home environment, nurturing and discipline behaviors, as well as parental perceived self-efficacy and perceived impact of own behaviors on child development.

In sum, changes in the level of parenting behaviors and beliefs account for about half the effect of NEP-I on our socioemotional index, but only a small fraction of the effect of NEP-I on vocabulary in the medium run. In the case of language, the impacts of NEP are likely to come through other unobserved channels, not discussed in this analysis. Finally, we do not find evidence that NEP influenced child outcomes through changes in the productivity of parenting behaviors and beliefs.

Table 9: Mediation analysis: Vocabulary Index

	(1)	(2)	(3)	(4)
	Base	+Behavioral Index	+Beliefs Index	+Significant Mediators
NEP_B	0.070	0.065	0.069	0.066
	(0.044)	(0.044)	(0.044)	(0.044)
NEP_I	0.099**	0.092**	0.096**	0.092**
	(0.045)	(0.045)	(0.045)	(0.045)
Caregiver Behavioral Index		0.049**		0.052**
		(0.020)		(0.022)
Caregiver Beliefs Index			0.016	-0.006
			(0.019)	(0.021)
Observations	2895	2895	2895	2895
% Indirect Effect		**7.9%**		**7.8%**
Confidence Intervals for the Joint Significance ($\beta_2^{I,l} * \pi_5^l$)				
Lower Bound		0.003%		0.003%
Upper Bound		12.86%		12.74%

Note: Each column reports estimates from a separate regression of the age-standardized child vocabulary index. Estimates control for child sex and gender, and for health center's fixed effects. Column 1 presents the ITT outcomes as a benchmark (Table 3). Columns 2-3 include one potential mediator at a time in the outcome equation. Column 4 include all significant mediators. The last rows report the mean and CI's of the total indirect effect in the child outcome that is attributable to intervention effects in mediators (as in equation (5)). Significance levels: *p<=10%, **p<=5%, ***p<=1%.

Table 10: Mediation analysis: Socioemotional Index

	(1)	(2)	(3)	(4)
	Base	+ Behavioral Index	+ Beliefs Index	+ Significant Mediators
NEP_B	0.032	0.013	0.024	0.012
	(0.047)	(0.045)	(0.046)	(0.045)
NEP_I	0.094**	0.061	0.064	0.050
	(0.047)	(0.045)	(0.046)	(0.045)
Behavioral Index		0.320***		0.247***
		(0.020)		(0.022)
Beliefs Index			0.267***	0.164***
			(0.020)	(0.021)
Observations	2492	2492	2492	2492
% Indirect Effect		**36.6%**	**31.4%**	**47.9%**
Confidence Intervals for the Joint Significance ($\beta_2^{I,l} * \pi_5^l$)				
Lower Bound		5.70%	5.03%	18.94%
Upper Bound		67.83%	58.34%	78.23%

Note: Each column reports estimates from a separate regression of the age-standardized child socioemotional index. Estimates control for child sex and gender, and for health center's fixed effects. Column 1 presents the ITT outcomes as a benchmark (Table 3). Columns 2-3 include one potential mediator at a time in the outcome equation. Column 4 include all significant mediators. The last rows report the mean and CI's of the total indirect effect in the child outcome that is attributable to intervention effects in mediators (as in equation (5). Significance levels: *p<=10%, **p<=5%, ***p<=1%.

7 Cost-Benefit Analysis

In this section, we perform the cost-benefit analysis (CBA) of NEP, and, more specifically, of NEP-I, the intervention arm that achieved statistically significant impacts on child developmental outcomes. We adopt a societal perspective where, in addition to program implementation costs, we account for future costs and benefits resulting from improvements in child development among program participants. Future costs include additional schooling costs associated with increased college attendance. Future benefits include net private gains in lifetime earnings, either directly by increasing wages or indirectly by increasing labor force participation (LFP), as well as societal gains associated with the reduction of crime and mental health problems.

Table 11 provides a detailed account of program implementation costs. We include (i) *labor costs*: facilitators were paid an hourly rate of $13 to conduct sessions, regardless of the program version, for an average of 13 (17) hours to administer 6 (8) group sessions of the NEP-B (NEP-I) program; (ii) *training costs*: facilitators assigned to NEP-B received 4 days of training and those assigned to NEP-I received 2 additional days of training. Since we do not have access to the training costs of the standard NEP-B program, we use the training costs of the NEP-I program to predict these training costs under conservative assumptions (see Appendix 7 for more details). To calculate

training costs per child, we assume that each facilitator can deliver the program without having to be re-trained to at least five groups, including the groups that were part of the evaluation sample. Finally, we also account for smaller costs related to the materials required to conduct the sessions. The overall cost per child of NEP-I (84.3 US$) is roughly 35% higher than NEP-B (62.2 US$). More details of how these costs are obtained are discussed in Appendix 7.

Table 11: Program Implementation Costs

	Unit Cost	Cost facilitator NEP-B	Cost facilitator NEP-I	NEP-B	NEP-I
	US$/hr-fac	US$/fac	US$/fac	US$/child	US$/child
Children per facilitator				6	6
Labor Costs				5	5
Facilitator hours (15 hrs.)	17.3	224.6	293.7	37.4	49.9
Materials					
Training Costs				10.9	14.6
Standard program NEP-B		413.6		13.8	
NEP-I (NEP-B + 2 extra sessions)		413.6	179.5		19.8
Total Cost per child				**62.2**	**84.3**

Notes: Each facilitator was paid 15 extra hours to deliver an average of 7 NEP-B sessions. Facilitators in the NEP-I arm were paid an additional 33% to deliver two extra sessions with parents and children. All costs used for calculations were originally in Chilean pesos and were converted to USD using exchange rate of 1 USD= 510 CLP as of July 2010, and converted into 2022 $USD adjusting by an inflation factor of 1.42 (3% annual).

To predict long-term costs and benefits resulting from program participation we need two inputs: (1) estimates of program impacts on each adult outcome (college attendance, earnings, LFP, etc.), which are unobserved because children in our sample are only 10-15 years old in 2022, and (2) measures of the present discounted value (PDV) of each of these lifetime costs and benefits. We use a 3% discount rate as a benchmark, in line with other cost-benefit analysis of similar programs in Latin America (Araujo et al. 2021, Berlinski and Schady 2015), and test the sensitivity of our results to a discount rate of 5%.

To estimate program impacts on adult outcomes, we need to map the medium-term program impacts on cognitive and socioemotional child development outcomes to adult outcomes. In this analysis, we use our IV estimates from Table 6 as they reflect the impact of participating of NEP.[18] One way to do so is by multiplying our medium-term impacts with estimates of the returns (in adult outcomes) to increases in children's cognitive and socioemotional development obtained from external longitudinal samples tracking child and adult outcomes. Such an approach has been used to estimate the long-term benefits of the STAR experiment in the US (Krueger 1999) and the impact of a package of ECD interventions in Nigeria (Carneiro et al. 2021). It relies on the strong

[18] However, in the Appendix 7 (Table A40), we show that our cost-benefit ratios using ITT estimates instead, remain sizable and our conclusions unchanged.

assumption that our medium-term estimates are good markers for long term impacts on adult outcomes, and that the assumed returns are applicable to our sample. Since such longitudinal data is not available for Chile, we estimate these returns using longitudinal data from the National Child Development Study (NCDS) in the UK. Specifically, we use the NCDS to estimate the gains in adult outcomes measured at age 43 including college attendance, earnings, LFP, crime, and depression, associated with a 1 SD increase in cognitive and socioemotional skills measured at age 7, controlling for covariates (for more details, see Table A39 in Appendix 7). For simplicity, we assume that the total program impact on a given adult outcome is the just sum of the impacts stemming from cognitive and socioemotional skill improvements.

Second, we need to calculate the PDV of lifetime costs and benefits for each adult outcome that we account for. For the societal cost associated with increased college attendance, we use the estimated cost of tertiary education per student in Chile in 2019 (Alessandri 2021) and compute its net present value. To estimate gains in lifetime earnings, either directly by improving wages or indirectly by increasing LFP, we obtain earnings and labor force participation rates profiles by age from a representative sample of individuals aged 25-64 from the 2019 *Encuesta Suplementaria de Ingresos* (Supplementary Income Survey). To estimate gains resulting from reductions in individuals' engagement in criminal activity, we use public data on the total cost of crime in Chile in 2014 (Saens 2015), and the total number of apprehended individuals in 2019 (CEAD 2019). Finally, lifetime gains from mental health improvements are obtained from three sources: (i) the public cost per person-year obtained from the 2021 national budget for mental health, (ii) the private (out-of-pocket) per person-year expenditures on mental health services obtained from the 2020 *Encuesta de Proteccion Social* (Social Protection Survey), and (iii) an estimate of labor costs savings from averting reduced labor participation associated with poor mental health (Ruiz-Tagle and Troncoso 2018).

Table 12 summarizes our findings. Columns 1-3 present the cost-benefit ratios that result from the estimated PDV of lifetime costs and benefits using the actual program costs from Table 11 and a discount rate of 3%. In columns 4-6 we present sensitivity analyses varying the program costs, assumptions about returns to adult outcomes, and the discount rate. Column 1 presents our benchmark cost-benefit ratios combining the actual program implementation costs, future college costs, and future gains in lifetime earnings. The PDV of the college cost is $334, which is the product of the predicted program impact on college attendance (5.8%) and the PDV of the cost of college attendance ($5,766). The PDV of gains in lifetime earnings among those who work is $8,352, which is the product of the PDV of lifetime earnings adjusted by labor force participation ($95,287), and the predicted program impact on earnings (8.8%). The resulting benchmark cost-benefit ratio is 20. If, in addition, we include gains in lifetime earnings resulting from increased labor force participation (Column 2) the cost-benefit ratio increases to 26. If we further include gains from crime reduction and mental health improvements, then the cost-benefit ratio increases to 30 (Column 3). The lion's share of societal benefits is represented by the lifetime gains in adult earnings.

Our estimated cost-benefit ratios are arguably large compared with those reported for similar programs that either use IV impacts to estimate long-term benefits or rely on ITT impacts of home visit interventions where compliance is less of an issue and thus the IV and ITT estimates are comparable (Berlinski and Schady 2015, Araujo et al. 2021). In the last three columns of Table

12, we perform a sensitivity analysis where we incrementally add more stringent assumptions. One potential concern is that our program costs might be understated, to the extent that the opportunity cost of infrastructure use and human capital costs beyond the extra hours paid to facilitators to conduct the sessions are not quantified. In Column 4, we show that quadrupling the program implementation costs translates into a reduction of the cost-benefit ratio from 30 to a still sizable ratio of 18.7. A second concern is that our assumed long-term returns might be too high: halving the estimated adult returns to changes in children's outcomes from NCDS data, further lowers the cost-benefit ratio to 12.4 (Column 5). If, in addition, we increase the discount rate to 5%, then the cost-benefit ratio further decreases to 7.3 but remains sizeable (Column 6).

Table 12: Cost-benefit Ratios

	Components of benefits			Sensitivity Analysis		
	(1) Earnings	(2) + LFP	(3) All costs and benefits	(4) 4 x Program costs	(5) + 50% of assumed returns	(6) + 5% discount rate
Costs						
Intervention costs	*$84*	*$84*	*$84*	*$337*	*$337*	*$337*
College attendance costs	$334	$334	*$334*	*$334*	*$167*	$118
Long-term societal costs	***$418***	***$418***	***$418***	***$671***	***$504***	***$455***
Benefits						
Gains in earnings	$8,352	$8,352	*$8,352*	*$8,352*	*$4,176*	*$2,171*
Gains due to increased LFP		$2,565	$2,565	*$2,565*	*$1,282*	*$663*
Gains in crime reduction			$1,236	*$1,236*	*$618*	*$371*
Gains in mental health			$386	$386	$193	*$110*
Long-term societal benefits	***$8,352***	***$10,917***	***$12,539***	***$12,539***	***$6,269***	***$3,315***
Cost-benefit Ratios	**20.0**	**26.1**	**30.0**	**18.7**	**12.4**	**7.3**

Note: The gain (or cost) in a specific adult outcome (e.g., earnings, labor force participation (LFP), crime reduction, mental health, college attendance) is the sum of the gains (or costs) induced by program improvements in language and socioemotional development. The predicted gain (or cost) in a specific adult outcome induced by program improvements in a specific child outcome (e.g., language) is obtained as the product between the IV intervention impact on the child outcome, the assumed correlation between the adult outcome and the child outcome, and the net benefit or cost per child, as described in the Appendix 7. All benefits and costs originally in Chilean pesos are converted to 2022 USD using the corresponding exchange rate to the year of the original data source and adjusted for inflation to January 2022. Present period total societal costs per child include direct program costs and the present value of long-term college attendance costs. Gains in adult outcomes are all in present values, where the discount rate is 3%.

8 Discussion

Our empirical analysis shows that a low-intensity, low-cost, and scaled-up parenting intervention, targeting vulnerable populations, and integrated with and delivered by the public health system, was able to generate medium-term impacts on child developmental outcomes, parenting behaviors, and parenting beliefs, almost three years after the end of the intervention period. We provide clear evidence that it is possible to effectively deliver a countrywide parenting program in a middle-

income country, with potentially meaningful human capital benefits for target children (who come primarily from low-income families). Successful delivery of this program requires a well-organized national health system with well-trained and motivated staff, which should be possible to encounter in other countries with similar or higher income levels.

The intensive version (NEP-I) of the program had larger impacts than the regular version (NEP-B), even though we can never reject the null hypothesis that the effect sizes are the same across intervention arms. We believe that these differences are the result of two factors. First, NEP-I facilitators received 2 additional days of training on top of the 4 days of training for the NEP-B version. These extra training sessions were designed by experts in early childhood development and were focused on improving responsive stimulation practices in caregiver-child interactions. The extra training enabled facilitators not only to learn the new sessions on play and language but also to integrate these new messages with the NEP-B curriculum resulting in a more comprehensive program overall. This integration was helped by the fact that the two sessions on parent-child interactions were placed in between the regular sessions and not only at the end.

Second, caregivers participating in NEP-I also attended more sessions than those in NEP-B. Theoretically, NEP-B consisted of 6 sessions and NEP-I added 2 more sessions on parent-child interactions. However, facilitators had the flexibility to program more sessions if they wanted to. While the median number of sessions attended in NEP-B was 6 and in NEP-I was 8, 23% of parents in NEP-I attended at least 10 sessions while this is true only for 5% in NEP-B, reflecting sharp differences in the distributions of parental attendance across the two programs. Moreover, the two additional sessions of caregiver-child interactions gave parents the opportunity to practice with their children and receive more personalized feedback across the whole program, enabling them to consolidate knowledge and behavioral change.

Some aspects of the curriculum design and results from our mediation analysis can provide some insights about why we find impacts in some outcomes and not in others. In terms of curriculum design, positive impacts of NEP-I on child vocabulary are likely to be explained by the larger emphasis given to responsive language and play in the two additional sessions of caregiver-child interactions included in this intervention arm. This also seems to be the case for socio-emotional outcomes, where the results are mostly driven by impacts in the Battelle measures (but not the CBCL), which capture adaptive behaviors that closely relate to the emphasis on responsive parent-child interactions that are the base of the NEP-I curriculum. Neither intervention was able to reduce maladaptive behaviors, the second component of our socioemotional index, but this may have been partly due to fieldwork problems concerning the collection of this data at endline: unfortunately, due to problems in the administration of the survey instrument at endline, the final sample for the CBCL scale was reduced by 1/3. Finally, we do not find impacts of NEP-I on executive functions, for which we cannot offer an explanation other than the program did not target these outcomes in the curriculum.

Effects on child outcomes appear to be partially mediated by changes in parental inputs, more so for impacts on socioemotional outcomes than for impacts on vocabulary. The mediation analysis presented in Tables 9 and 10 suggests that improvements in the behavioral index explain only 8% of the impact of the intervention on child vocabulary, and about 37% of the intervention effect on child socioemotional index, respectively. Interestingly, a composite index of parenting beliefs also accounts for 31% of the impact of NEP-I on socio-emotional outcomes when included separately. When both indices are jointly included, the mediation analysis accounts for about 48% of the

impact of NEP-I on socio-emotional outcomes, and the offer of the program is no longer a statistically significant predictor of outcomes. In sum, our analyses suggest that changes in behaviors and beliefs account for about half of the effect of NEP-I on child socioemotional development, and changes in behaviors only account for a small fraction of the effect of NEP-I on vocabulary in the medium run.

Finally, we found that the total implementation costs were $62 per child for the standard NEP-B program and $84 for NEP-I. These costs are significantly smaller than other high-intensity and longer parenting interventions such as the home visiting programs of Colombia and Peru, with estimated annual costs per child of US $500 and US $300, respectively (Araujo et al. 2021). Coupled with the fact that our medium-term ITT impacts are similar or higher than those from these interventions (0.1 SD in Peru, and impacts in Colombia faded out two years after the end of the program), NEP becomes a highly cost-effective program in comparison. Adopting a societal perspective to include future costs and benefits associated with participation in NEP, we also found that the long-term economic returns of NEP could be extremely high, with cost-benefit ratios ranging between 20 and 30. Even if program costs were severely underestimated, and under extremely conservative assumptions about the discount rate and the long-term returns to children's cognitive and socioemotional development, the program's cost-benefit ratios remain very high, which indicates that such a program would be a good candidate to scale-up in other similar settings. Comparatively, the NEP cost-benefit ratios are higher than those estimated for the Peruvian and Colombian home visiting programs (5.4 and 4.6, respectively), much higher than those reported for the US Nurse Family Partnership Program (range 1.5-5) (Glazner et al. 2004, Heckman et al. 2017), and higher than other five home visitation programs in Latin America (Berlinski and Schady 2015, Boo, Palloni, and Urzua 2014, Walker et al. 2015).

Our findings are important for two reasons. First, because there is very little evidence, from both developed and low- and middle-income countries, about the ability of ECD interventions fully integrated in the health system and focusing on parental behavioral change to achieve sustained impacts on child outcomes that outlast the duration of the program. Our literature review in the introduction shows that very rarely successful ECD interventions targeting parenting in the short-term are followed up over time and results are mixed. In developed countries, despite multiple systematic reviews pointing out to the short-term effects of low-intensity, group-based, and parents-only programs such as Triple-P (Sanders et al. 2014), all these studies are small-scale efficacy trials, find impacts on children's socioemotional but not cognitive development, and only a couple of studies find sustained impacts over time (Kim et al. 2018). In LMIC settings, the few parenting interventions that include longer-term follow-ups (almost all of them delivered through individual home visits) find that early program benefits in child development tend to fade out over time (Jeong, Pitchik, and Fink 2021). Recent evidence from very intensive home visiting and group-based programs extended for at least two years such as the Preparing for Life program in Ireland (Doyle 2020) and the Nurse Family Partnership in the US (Heckman et al. 2017) shows that these programs were more likely to attain sustained impacts over time. However, all these programs are much more intensive and expensive than NEP. The cost per family per year of the PFL program was US$2,250 and of the NFP program roughly twice as much, which would make them unaffordable for a country like Chile.

Second, the size of the estimated ITT program impacts three years later (about 0.1 SD), combined with its low program costs, make NEP remarkably cost-effective in comparison with the most successful (yet much costlier) programs highlighted above. Pakistan's LHW program (Yousafzai

et al. 2014), which integrated two years of monthly home visits within the existing health services, achieved short-term impacts (measured immediately after the end of the intervention) on cognitive and socioemotional outcomes in the range 0.6-0.7 SD. However, a subsequent evaluation measuring medium term impacts two years after the end of the intervention finds that these effects substantially declined to 0.1-0.3 SD in cognition and to 0.2 SD in socio-emotional outcomes (Yousafzai et al. 2016)Click or tap here to enter text., with magnitudes that are more comparable to our study within a similar time window post-intervention. The NFP, which consisted of two years of home visits, had positive short-term impacts at 24 months in the quality of the home environment and behavioral problems among girls. Follow-up evaluations showed that the intervention also improved cognition (0.12-0.27 SD) and prosocial skills for girls (0.36 SD) at age 6, language development at age 9 (0.24 SD), and reduced internalizing problems for boys at age 12 (Olds et al. 2004, Heckman et al. 2017). The PFL program, which offered continued support to families from pregnancy until age 5 through a combination of individual home visits and group meetings, significantly improved self-reported measures of child cognition (0.22-0.42 SD), behavioral problems (0.18-0.24 SD) and pro-social behaviors (0.35 D) measured at 24, 36 and 48 months, as well as direct assessments of child cognition by 0.77 SD at the end of the program. Even though the medium-term impacts of these programs are generally larger than ours, the fact that NEP-I can achieve sustained impacts with a far less intensive intervention and at a cost that is orders of magnitude lower, reflects the high relative cost-effectiveness of NEP-I and its scalability potential.

Finally, our IV impacts (of about 0.4-0.5 SD) are more comparable with the medium-term impacts of the other more intensive programs, such as NFP and PFL and the Triple P programs. That said, we acknowledge that , our IV results need to be interpreted with caution. Given that the compliance rates among eligible families are below the 50% mark, it is possible that our IV effects for the sub-sample of compliers are larger than the average treatment effect among all eligible families. Our tests for selection on unobservables performed in section 6.2 seem to support this hypothesis. We find consistent evidence that program take-up in NEP-I was driven by families with children exhibiting lower child development scores and caregivers with worse parenting skills and beliefs. This negative selection means that families that needed the program the most, and perhaps had the highest returns from participating in the program were those who actually attended. It is possible that expanding this program to a wider population might not result in impacts of a comparable magnitude.

9 Conclusion

There is a large consensus across disciplines on the importance of high-quality interventions during the early years, a period in which critical cognitive and socio-emotional development processes are consolidated, with long-term implications for adulthood. Human capital investments during early childhood are not only important on the grounds of efficiency, given that earlier interventions have larger returns in the long-term, but also from the point of view of equity, as early childhood interventions are likely to reduce socio-economic gaps and the intergenerational transmission of poverty. Parents and caregivers play a key role in home stimulation during the early years, which is fundamental for healthy child development and is crucial to close early socioeconomic gaps in skills development.

This paper studies the medium-term results of a large-scale parenting program in Chile. The intervention, known as *Nadie es Perfecto* or NEP, provides information and support to parents and caregivers of the poorest and more disadvantaged groups, using a semi-structured curriculum, where trained and certified facilitators who encourage group discussions about parental needs and concerns. The curriculum is based on experiential learning. Parents and caregivers share and learn from other parents' experiences and discuss the challenges of parenting that prevent the adoption of new strategies at home. The main objective of NEP is to change parental beliefs about their role as parents, and facilitate the adoption of positive practices to foster a better parent-child interaction which, in turn, translates into better child developmental outcomes.

Our results show sustained effects of the program on parenting beliefs, practices and child outcomes three years after the intervention ended. The impacts of the offer of the program show a significant positive effect of 0.1 SD in vocabulary and of 0.09 SD in socio-emotional development. These treatment effects in child outcomes are mirrored by a sustained improvement of 0.11 SD in parenting behaviors including cognitive stimulation, nurturing, and discipline practices, as well as by a sustained improvement of 0.11 SD in parenting beliefs including perceived self-efficacy, perceived impact of own behavior on their children's development, and perceived social support. The offer of the program did not have any impact on parental psychological well-being. Accounting for an effective attendance to group sessions that ranged between 25% in NEP-B to 31% in NEP-I, implies a substantial improvement of 0.43 SD in language development, 0.38 SD in social development, 0.46 SD in parenting behaviors and 0.46 SD in parenting beliefs among participants. However, the estimated treatment effects accounting for participation are only applicable to the sub-population of compliers and cannot be extrapolated to the whole population.

Our results suggest that NEP seems to operate by improving parenting strategies with children and by changing parenting beliefs. Results from a mediation analysis suggest that both changes in parenting behaviors and beliefs greatly mediate NEP impacts on socio-emotional outcomes as they jointly explain about 48% of the total effect of the intervention in this outcome. However, in the case of receptive language only parenting behaviors play a modest mediating role as they explain only about 8% of the total effect of NEP in this outcome.

Our findings also show that NEP is a highly cost-effective program when compared with other parenting interventions from similar settings, and the long-term benefits of scaling up this program would vastly outweigh current and future program's costs, even under the most conservative scenarios of the assumptions used to predict future costs and benefits.

We believe our paper makes an important contribution to the literature on early human capital formation by providing evidence about the effectiveness of a parenting program that exhibits sustained changes in child developmental outcomes, as well as in parenting behaviors and beliefs. The strength of the program revolves around a semi-structured curriculum that adapts to parental interests and needs, its low cost and intensity, and its suitability to be integrated within existing health platforms and services. Important future research avenues include a better understanding of the role of local social networks in sustaining impacts, as well as collecting new survey data combined with rich administrative data available to document the program's long-term impacts on education and health outcomes.

References

Abidin, Richard R. 1990. *Parenting stress index-short form*: Pediatric psychology press Charlottesville, VA.

Aboud, Frances E, and Aisha K Yousafzai. 2015. "Global health and development in early childhood." *Annu Rev Psychol* 66 (1):433-457.

Achenbach, Thomas M, and Thomas M Ruffle. 2000. "The Child Behavior Checklist and related forms for assessing behavioral/emotional problems and competencies." *Pediatrics in review* 21 (8):265-271.

Alessandri, Francisco. 2021. Comparación Gasto Publico por Nivel Educativo. In *Policy Brief*: AcciónEducar.

Andrew, A., O. Attanasio, E. Fitzsimons, S. Grantham-McGregor, C. Meghir, and M. Rubio-Codina. 2018. "Impacts 2 years after a scalable early childhood development intervention to increase psychosocial stimulation in the home: A follow-up of a cluster randomised controlled trial in Colombia." *PLoS Med* 15 (4):e1002556. doi: 10.1371/journal.pmed.1002556.

Angrist, Joshua, Eric Bettinger, and Michael Kremer. 2006. "Long-Term Educational Consequences of Secondary School Vouchers: Evidence from Administrative Records in Colombia." *American Economic Review* 96:847-862. doi: 10.1257/aer.96.3.847.

Araujo, M Caridad, Marta Dormal, Sally Grantham-McGregor, Fabiola Lazarte, Marta Rubio-Codina, and Norbert Schady. 2021. "Home visiting at scale and child development." *Journal of Public Economics Plus* 2:100003.

Attanasio, O. P., C. Fernandez, E. O. Fitzsimons, S. M. Grantham-McGregor, C. Meghir, and M. Rubio-Codina. 2014. "Using the infrastructure of a conditional cash transfer program to deliver a scalable integrated early child development program in Colombia: cluster randomized controlled trial." *BMJ* 349:g5785. doi: 10.1136/bmj.g5785.

Attanasio, Orazio, Flávio Cunha, and Pamela Jervis. 2019. Subjective parental beliefs. their measurement and role. National Bureau of Economic Research.

Bandura, A. 1986. *Social Foundations of Thought and Action: A Social Cognitive Theory*: Prentice-Hall.

Bandura, Albert. 1995. *Self-Efficacy in Changing Societies*. Cambridge: Cambridge University Press.

Baranov, Victoria, Sonia Bhalotra, Pietro Biroli, and Joanna Maselko. 2020. "Maternal depression, women's empowerment, and parental investment: evidence from a randomized controlled trial." *American economic review* 110 (3):824-59.

Barlow, Jane, Hanna Bergman, Hege Kornør, Yinghui Wei, and Cathy Bennett. 2016. "Group-based parent training programmes for improving emotional and behavioural adjustment in young children." *Cochrane database of systematic reviews* (8).

Baumrind, Diana. 1968. "Authoritarian vs. authoritative parental control." *Adolescence* 3 (11):255.

Berlinski, Samuel, and Norbert Schady. 2015. "More bang for the buck: investing in early childhood development." In *The early years*, 149-178. Springer.

Black, Dan, Joonhwi Joo, Robert LaLonde, Jeffrey Smith, and Evan Taylor. 2020. Simple Tests for Selection: Learning More from Instrumental Variables. Human Capital and Economic Opportunity Working Group.

Blair, Clancy, and Rachel Peters Razza. 2007. "Relating effortful control, executive function, and false belief understanding to emerging math and literacy ability in kindergarten." *Child development* 78 (2):647-663.
Bock, R Darrell, and Robert D Gibbons. 2021. *Item response theory*: John Wiley & Sons.
Boivin, Michel, Daniel Pérusse, Ginette Dionne, Valérie Saysset, Mark Zoccolillo, George M Tarabulsy, Nathalie Tremblay, and Richard E Tremblay. 2005. "The genetic-environmental etiology of parents' perceptions and self-assessed behaviours toward their 5-month-old infants in a large twin and singleton sample." *Journal of Child Psychology and Psychiatry* 46 (6):612-630.
Boneva, Teodora, and Christopher Rauh. 2018. "Parental beliefs about returns to educational investments—the later the better?" *Journal of the European Economic Association* 16 (6):1669-1711.
Boo, Florencia Lopez, Giordano Palloni, and Sergio Urzua. 2014. "Cost–benefit analysis of a micronutrient supplementation and early childhood stimulation program in Nicaragua." *Annals of the New York Academy of Sciences* 1308 (1):139-148.
Bradley, Robert H, and Bettye M Caldwell. 1984. "The HOME Inventory and family demographics." *Developmental Psychology* 20 (2):315.
Briscoe, Ciara, and Frances Aboud. 2012. "Behaviour change communication targeting four health behaviours in developing countries: a review of change techniques." *Social science & medicine* 75 (4):612-621.
Britto, Pia R, Stephen J Lye, Kerrie Proulx, Aisha K Yousafzai, Stephen G Matthews, Tyler Vaivada, Rafael Perez-Escamilla, Nirmala Rao, Patrick Ip, and Lia CH Fernald. 2017. "Nurturing care: promoting early childhood development." *The Lancet* 389 (10064):91-102.
Camehl, Georg F, Christa Katharina Spiess, and Kurt Hahlweg. 2020. "The effects of a parenting program on maternal well-being: Evidence from a Randomized Controlled Trial." *The BE Journal of Economic Analysis & Policy* 20 (4).
Campbell, Frances A, and Craig T Ramey. 1994. "Effects of early intervention on intellectual and academic achievement: a follow-up study of children from low-income families." *Child development* 65 (2):684-698.
Carneiro, Pedro, Lucy Kraftman, Giacomo Mason, Lucie Moore, Imran Rasul, and Molly Scott. 2021. "The impacts of a multifaceted prenatal intervention on human capital accumulation in early life." *American Economic Review* 111 (8):2506-49.
CEAD. 2019. Estatisticas Delictuales. Ministerio del Interior y Seguridad Publica: Centro de Estudios y Analysis del Delito.
Cunha, Flávio, Irma Elo, and Jennifer Culhane. 2013. Eliciting maternal expectations about the technology of cognitive skill formation. National Bureau of Economic Research.
Cunha, Flávio, Irma Elo, and Jennifer Culhane. 2022. "Maternal subjective expectations about the technology of skill formation predict investments in children one year later." *Journal of Econometrics* 231 (1):3-32. doi: https://doi.org/10.1016/j.jeconom.2020.07.044.
Cunha, Flavio, and James J Heckman. 2008. "Formulating, identifying and estimating the technology of cognitive and noncognitive skill formation." *Journal of human resources* 43 (4):738-782.
Cunha, Flavio, James J Heckman, and Susanne M Schennach. 2010. "Estimating the technology of cognitive and noncognitive skill formation." *Econometrica* 78 (3):883-931.

Cutrona, Carolyn E, and Beth R Troutman. 1986. "Social support, infant temperament, and parenting self-efficacy: A mediational model of postpartum depression." *Child development*:1507-1518.
Davis, Jonathan MV, Jonathan Guryan, Kelly Hallberg, and Jens Ludwig. 2017. The economics of scale-up. National Bureau of Economic Research.
Doepke, Matthias, and Fabrizio Zilibotti. 2017. "Parenting with style: Altruism and paternalism in intergenerational preference transmission." *Econometrica* 85 (5):1331-1371.
Doyle, Orla. 2020. "The first 2,000 days and child skills." *Journal of Political Economy* 128 (6):2067-2122.
Echeverría, M, M Herrera, and J Segure. 2002. "TEVI-R, Test de Vocabulario en Imágenes [TEVI-R Picture Vocabulary Test]." *Concepción, Chile: Publicaciones Universidad de Concepción*.
Engle, Patrice L, Lia CH Fernald, Harold Alderman, Jere Behrman, Chloe O'Gara, Aisha Yousafzai, Meena Cabral De Mello, Melissa Hidrobo, Nurper Ulkuer, and Ilgi Ertem. 2011. "Strategies for reducing inequalities and improving developmental outcomes for young children in low-income and middle-income countries." *The Lancet* 378 (9799):1339-1353.
Fernald, Lia CH, Patricia Kariger, Melissa Hidrobo, and Paul J Gertler. 2012. "Socioeconomic gradients in child development in very young children: Evidence from India, Indonesia, Peru, and Senegal." *Proceedings of the National Academy of Sciences* 109 (supplement_2):17273-17280.
Fox, Robert A. 1994. *Parent behavior checklist*.
Furlong, M., S. McGilloway, T. Bywater, J. Hutchings, S. M. Smith, and M. Donnelly. 2012. "Behavioural and cognitive-behavioural group-based parenting programmes for early-onset conduct problems in children aged 3 to 12 years." *Cochrane Database Syst Rev* (2):Cd008225. doi: 10.1002/14651858.CD008225.pub2.
Gertler, Paul, James Heckman, Rodrigo Pinto, Arianna Zanolini, Christel Vermeersch, Susan Walker, Susan M Chang, and Sally Grantham-McGregor. 2014. "Labor market returns to an early childhood stimulation intervention in Jamaica." *Science* 344 (6187):998-1001.
Glazner, Judith, Jessica Bondy, Dennis Luckey, and David Olds. 2004. "Final Report To The Administration For Children And Families: Effect of the Nurse Family Partnership on Government Expenditures for Vulnerable First-Time Mothers And their Children in Elmira, New York, Memphis, Tennessee, and Denver, Colorado (# 90XP0017)[Internet]." *Research and Evaluation*.
Goldberg, Lewis R. 1993. "The structure of phenotypic personality traits." *American psychologist* 48 (1):26.
Grantham-McGregor, Sally, Akanksha Adya, Orazio Attanasio, Britta Augsburg, Jere Behrman, Bet Caeyers, Monimalika Day, Pamela Jervis, Reema Kochar, and Prerna Makkar. 2020. "Group sessions or home visits for early childhood development in India: a cluster RCT." *Pediatrics* 146 (6).
Grantham-McGregor, Sally M, Christine A Powell, Susan P Walker, and John H Himes. 1991. "Nutritional supplementation, psychosocial stimulation, and mental development of stunted children: the Jamaican Study." *The Lancet* 338 (8758):1-5.
Hamadani, Jena D, Fahmida Tofail, Afroza Hilaly, Syed N Huda, Patrice Engle, and Sally M Grantham-McGregor. 2010. "Use of family care indicators and their relationship with

child development in Bangladesh." *Journal of health, population, and nutrition* 28 (1):23.
Heckman, James J, Margaret L Holland, Kevin K Makino, Rodrigo Pinto, and Maria Rosales-Rueda. 2017. An analysis of the memphis nurse-family partnership program. National Bureau of Economic Research.
Heckman, James, and Dimitriy V. Masterov. 2007. "The Productivity Argument for Investing in Young Children *." *Review of Agricultural Economics* 29 (3):446-493.
Heckman, James, Seong Hyeok Moon, Rodrigo Pinto, Peter Savelyev, and Adam Yavitz. 2010. "Analyzing social experiments as implemented: A reexamination of the evidence from the HighScope Perry Preschool Program." *Quantitative economics* 1 (1):1-46.
Heming, G, PA Cowan, and CP Cowan. 1990. "Ideas about parenting." *Handbook of family measurement techniques*:362-363.
Jeong, Joshua, Emily E Franchett, Clariana V Ramos de Oliveira, Karima Rehmani, and Aisha K Yousafzai. 2021. "Parenting interventions to promote early child development in the first three years of life: A global systematic review and meta-analysis." *PLoS medicine* 18 (5):e1003602.
Jeong, Joshua, Helen O Pitchik, and Günther Fink. 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 (3):e004067.
John, Oliver P, Eileen M Donahue, and Robert L Kentle. 1991. "Big five inventory." *Journal of Personality and Social Psychology*.
Justino, Patricia, Marinella Leone, Pierfrancesco Rolla, Monique Abimpaye, Caroline Dusabe, Diane Uwamahoro, and Richard Germond. 2020. Improving parenting practices for early child development: Experimental evidence from Rwanda. World Institute for Development Economic Research (UNU-WIDER).
Kagitcibasi, Cigdem, Diane Sunar, Sevda Bekman, Nazli Baydar, and Zeynep Cemalcilar. 2009. "Continuing effects of early enrichment in adult life: The Turkish Early Enrichment Project 22 years later." *Journal of Applied Developmental Psychology* 30 (6):764-779.
Kim, Jun Hyung, Wolfgang Schulz, Tanja Zimmermann, and Kurt Hahlweg. 2018. "Parent–child interactions and child outcomes: Evidence from randomized intervention." *Labour Economics* 54:152-171.
Knight, Robert G, Sheila Williams, Rob McGee, and Susan Olaman. 1997. "Psychometric properties of the Centre for Epidemiologic Studies Depression Scale (CES-D) in a sample of women in middle life." *Behaviour research and therapy* 35 (4):373-380.
Kolb, David A. 2014. *Experiential learning: Experience as the source of learning and development*: FT press.
Krueger, Alan B. 1999. "Experimental estimates of education production functions." *The quarterly journal of economics* 114 (2):497-532.
Lee, David S. 2009. "Training, wages, and sample selection: Estimating sharp bounds on treatment effects." *The Review of Economic Studies* 76 (3):1071-1102.
Lu, Irene RR, D Roland Thomas, and Bruno D Zumbo. 2005. "Embedding IRT in structural equation models: A comparison with regression based on IRT scores." *Structural Equation Modeling* 12 (2):263-277.
Luoto, Jill E, Italo Lopez Garcia, Frances E Aboud, Daisy R Singla, Lia CH Fernald, Helen O Pitchik, Uzaib Y Saya, Ronald Otieno, and Edith Alu. 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 (3):e309-e319.
Maccoby, E. E., A. J. Kahn, and B. A. Everett. 1983. "The role of psychological research in the formation of policies affecting children." *Am Psychol* 38 (1):80-4. doi: 10.1037//0003-066x.38.1.80.
Ohan, Jeneva L, Debbie W Leung, and Charlotte Johnston. 2000. "The Parenting Sense of Competence scale: Evidence of a stable factor structure and validity." *Canadian Journal of Behavioural Science/Revue canadienne des sciences du comportement* 32 (4):251.
Olds, David L, Harriet Kitzman, Robert Cole, JoAnn Robinson, Kimberly Sidora, Dennis W Luckey, Charles R Henderson Jr, Carole Hanks, Jessica Bondy, and John Holmberg. 2004. "Effects of nurse home-visiting on maternal life course and child development: age 6 follow-up results of a randomized trial." *Pediatrics* 114 (6):1550-1559.
Ringwalt, Sharon. 2008. "Developmental Screening and Assessment Instruments with an Emphasis on Social and Emotional Development for Young Children Ages Birth through Five." *National Early Childhood Technical Assistance Center (NECTAC)*.
Robins, James M, Andrea Rotnitzky, and Lue Ping Zhao. 1994. "Estimation of regression coefficients when some regressors are not always observed." *Journal of the American statistical Association* 89 (427):846-866.
Roid, Gale H, and Lucy J Miller. 1997. "Leiter international performance scale-revised (Leiter-R)." *Wood Dale, IL: Stoelting* 10.
Romano, Joseph P, and Michael Wolf. 2005. "Stepwise multiple testing as formalized data snooping." *Econometrica* 73 (4):1237-1282.
Ruiz-Tagle, Jaime, and Pablo Troncoso. 2018. Labor Cost of Mental Health: Evidence from Chile. In *Working Papers wp468*, edited by Department of Economics University of Chile
Saens, Rodrigo. 2015. ¿Cuánto cuesta el delito en Chile? edited by Chile Universidad de Talca.
Sanders, Matthew R. 2012. "Development, evaluation, and multinational dissemination of the Triple P-Positive Parenting Program." *Annual review of clinical psychology* 8:345-379.
Sanders, Matthew R, James N Kirby, Cassandra L Tellegen, and Jamin J Day. 2014. "The Triple P-Positive Parenting Program: A systematic review and meta-analysis of a multi-level system of parenting support." *Clinical psychology review* 34 (4):337-357.
Sylvia, Sean, Nele Warrinnier, Renfu Luo, Ai Yue, Orazio Attanasio, Alexis Medina, and Scott Rozelle. 2020. "From Quantity to Quality: Delivering a Home-Based Parenting Intervention Through China's Family Planning Cadres." *The Economic Journal* 131 (635):1365-1400. doi: 10.1093/ej/ueaa114.
Todd, Petra E, and Kenneth I Wolpin. 2007. "The production of cognitive achievement in children: Home, school, and racial test score gaps." *Journal of Human capital* 1 (1):91-136.
Waldrop, Julee, Maureen Baker, Rebecca Salomon, and Elizabeth Moreton. 2021. "Parenting Interventions and Secondary Outcomes Related to Maternal Mental Health: A Systematic Review." *Maternal and Child Health Journal* 25 (6):870-880.
Walker, Susan P, Christine Powell, Susan M Chang, Helen Baker-Henningham, Sally Grantham-McGregor, Marcos Vera-Hernandez, and Florencia López-Boo. 2015. Delivering parenting interventions through health services in the Caribbean: Impact, acceptability and costs. IDB Working Paper Series.

Webster-Stratton, Carolyn. 2001. "The incredible years: Parents, teachers, and children training series." *Residential treatment for children & youth* 18 (3):31-45.

Yousafzai, A. K., M. A. Rasheed, A. Rizvi, R. Armstrong, and Z. A. Bhutta. 2014. "Effect of integrated responsive stimulation and nutrition interventions in the Lady Health Worker programme in Pakistan on child development, growth, and health outcomes: a cluster-randomised factorial effectiveness trial." *Lancet* 384 (9950):1282-93. doi: 10.1016/s0140-6736(14)60455-4.

Yousafzai, Aisha K, Jelena Obradović, Muneera A Rasheed, Arjumand Rizvi, Ximena A Portilla, Nicole Tirado-Strayer, Saima Siyal, and Uzma Memon. 2016. "Effects of responsive stimulation and nutrition interventions on children's development and growth at age 4 years in a disadvantaged population in Pakistan: a longitudinal follow-up of a cluster-randomised factorial effectiveness trial." *The Lancet Global Health* 4 (8):e548-e558.

Zelazo, Philip David. 2006. "The Dimensional Change Card Sort (DCCS): A method of assessing executive function in children." *Nature protocols* 1 (1):297-301.

Appendix (supplemental material not for publication)

Appendix 0: Program details and study design

Health center eligibility: In order to implement the intervention, eligible health centers had to identify available rooms to carry out the group sessions, as well as hire facilitator's assistants to provide childcare for caregivers attending the session who were not able to leave children at home. The Chile Crece Contigo system provided health centers the funding to cover childcare expenses and some of the facilitators' time allocated to prepare and deliver the sessions. The health centers were expected to embed the facilitators' time allocated to the NEP sessions within their regular tasks.

Training of facilitators: Training into the NEP-B curriculum followed a Training-of-Trainers model. In a first stage, 32 master trainers (one per health region or "servicio de salud") were trained directly by an international lead trainer from NEP Canada and it comprised two parts: a) a training module to become group facilitator consisting of 40 theoretical hours and the practical implementation of two NEP-B sessions; and b) a training module to receive formal certification as a NEP trainer consisting of 40 training hours and the practical replications of three group facilitator training sessions. All master trainers had a university degree and significant experience in the health sector as nurses, psychologists, occupational therapists, and social workers. In a second stage, master trainers trained more than 1,700 facilitators between the end of 2009 and the beginning of 2010 to deliver NEP-B across more than 300 municipalities in the whole country. These training sessions lasted 32 hours (4 consecutive days) and included both theoretical and practical implementation of sessions.

For the two additional sessions on parent-child interactions designed by the team of psychologists from the program *Juguemos con Nuestros Hijos* at *U. Catolica*, NEP-I facilitators received 2 additional days of training. These extra training sessions were of very high quality and were focused on improving responsive play and communication practices in guided caregiver-child interactions. This enabled facilitators not only to learn the new sessions on play and language but also to integrate these new messages with the NEP-B curriculum resulting in a more comprehensive program overall.

The Training-of-Trainers model used for NEP is highly scalable, and over the years, the 32 master trainers have continued to train new NEP facilitators, and sometimes new master trainers, to replace those who left their position.

Recruitment of facilitators: Certified trainers carried out formal training of group facilitators over a period of a year. The number of facilitators per health center to be trained was calculated based on the size of the target population. A training call was done for health professionals within the health centers, preferably among those who had previous experience in early childhood development. NEP relies on facilitators with a university degree who, with the appropriate training, can deliver a curriculum that is highly flexible in order to accommodate family needs.

Monitoring and Supervision: The program monitoring and supervision are carried out by master trainers, as part of their duties. These activities include regular contacts over the phone with

facilitators, as well as visits to the different health centers belonging to their health region to observe the NEP sessions and give facilitators feedback. In addition, facilitators are required to upload the logs of the sessions to the centralized online system that include information on session attendance, topics discussed, and problems found in delivering the session. These records are compiled and analyzed by a central team at the Ministry of Health in order to provide master trainers with feedback about the quality of implementation in the health centers under their supervision.

Program dissemination and target population: Regular health checks represent the most frequent point of contact between the health centers and families, where caregivers are invited to participate in the program. NEP also promotes the program through posters and leaflets in waiting rooms at health centers. Facilitators also recruited participants through presentation sessions at local preschools for recruitment.
Exclusion criteria included household with multiple risks such as detected domestic violence, severe mental health problems (such as severe depression or personality disorders), or severe child developmental delays that require clinical attention. Households with these multiple adversities are referred to other health specialists within the health center to receive more intensive and personalized services. For all households meeting the inclusion criteria, NEP is conceived as a universal program for all caregivers who are interested in improving their parenting skills, as well as a more targeted program for households who are perceived as vulnerable according to a) a screening psychosocial evaluation scale administered in the regular health check-ups, or b) the assessment of risk factors by local health professionals including single parents, adolescent parents, low income/low literacy parents, or parents with kids with moderate behavioral problems.

Appendix 1: Baseline Sample Characteristics and Outcomes

Table A1 shows means and standard deviations of children's performance in receptive and expressive language development measured at baseline using the PLS-IV scale. There are no significant differences across treatment arms either when we use global T scores, or when we use the T scores for the receptive and expressive language sub-dimensions. Using the global T scores to diagnose developmental delays, we find that 16.7% of our sample between 3 months and 5 years old are diagnosed with some degree of delay and that 5.8% of children are diagnosed with a clinical delay.

Table A1: Baseline Balance: Child Receptive and Expressive Language

	(1) Control	(2) NEP-B	(3) NEP-I	t-test p-value	t-test p-value
Language score (PLSIV)	Mean/SE	Mean/SE	Mean/SE	(1)-(2)	(1)-(3)
Global score	99.453	100.35	99.553	0.235	0.895
	(0.529)	(0.539)	(0.544)		
Receptive Language score	101.146	102.225	101.157	0.161	0.988
	(0.544)	(0.545)	(0.552)		

Expressive Language score	97.574	98.063	97.732	0.492	0.826
	(0.491)	(0.516)	(0.524)		
Diagnosis (Based on Global score)					
Clinical range (%)	0.058	0.047	0.07	0.268	0.266
Risk (%)	0.114	0.114	0.097	0.983	0.211
Normal (%)	0.828	0.839	0.833	0.518	0.763
Observations	1089	1060	1049		
F-test of joint significance (p-value)	0.719	0.388			

Note: T-tests report comparisons between the control arm against NEP-B and NEP-I. Significance levels: *p<=10%,

Table A2 describes child internalizing and externalizing behavioral problems reported by caregivers using the Child Behavior Checklist scale. The survey is administered to mothers of all children between 18 months and 5 years old. In our sample, 28.5% of children show some mild or severe level of alteration (27.3% internalizing and 19.5% externalizing). There are no significant differences in scores across groups for any sub-dimension, and while there is a marginally significant difference between the Control group and NEP Basic in the percentage of children with moderate risk, the joint test across variables suggests a very low risk of sample imbalance.

Table A2: Baseline Balance: Child Maladaptive Behavior

	(1) Control	(2) NEP-B	(3) NEP-I	t-test p-value	t-test p-value
Maladaptive Behavior, CBCL test	Mean/SE	Mean/SE	Mean/SE	(1)-(2)	(1)-(3)
T score, Global	56.828	57.094	56.316	0.676	0.430
	(0.451)	(0.448)	(0.468)		
T score, Internalization	56.130	56.290	55.939	0.802	0.765
	(0.449)	(0.450)	(0.456)		
T score, Externalization	54.990	55.299	54.595	0.595	0.508
	(0.412)	(0.411)	(0.433)		
Diagnosis (based on Global score)					
Clinical range (%)	0.155	0.135	0.143	0.270	0.518
Risk (%)	0.125	0.156	0.138	0.083*	0.466
Normal (%)	0.720	0.709	0.719	0.635	0.972
Observations	774	769	754		
F-test of joint significance (p-value)				0.516	0.681

Note: T-tests report comparisons between the control arm against NEP-B and NEP-I. Significance levels: *p<=10%, **p<=5%, ***p<=1%. F-test for the joint significance across all variables is reported at the bottom.

Table A3 shows the Dimensional Card Sort scale (DCCS) measure of executive function performance in children older than 24 months old. In the test, if the child does not pass the first

stage, she cannot be evaluated, which means that her performance is too low to be measured by the scale. If the child passes the first stage, she is evaluated as "Normal" if she completes the task, or "Altered" if she leaves the task incomplete. For example, the table shows that the proportion of children with "Altered" results out of those who passed the first stage is about 19.7% for children in the 36-47 months range, and 17.6% for children in the 48-59 months range and 11.4% for older children. We did not find any significant differences in the diagnostic across groups, except for the percentage of children with altered scores in the age groups 24-25 and 60-72 months, and the percentage of children that fail to pass the pre-change in the age group 60-72 months. Once again, the sample is fairly balanced across the three treatment arms.

Table A3: Baseline Balance: Executive function performance

	(1)	(2)	(3)	t-test	t-test
	Control	NEP-B	NEP-I	(1)-(2)	(1)-(3)
Executive Function (DCCS)	Mean/SE	Mean/SE	Mean/SE	(1)-(2)	(1)-(3)
24-35 months					
Score	6.237	6.126	6.064	0.764	0.646
	(0.265)	(0.255)	(0.267)		
Fail to pass pre-stage (%)	0.711	0.722	0.706	0.800	0.920
Altered (%)	0.137	0.078	0.123	0.051	0.674
Normal (%)	0.153	0.200	0.172	0.208	0.612
Observations	190	230	204		
36-47 months					
Score	9.162	8.939	9.039	0.471	0.682
	(0.205)	(0.232)	(0.218)		
Fail to pass pre-stage (%)	0.332	0.356	0.305	0.619	0.560
Altered (%)	0.179	0.194	0.202	0.700	0.554
Normal (%)	0.489	0.450	0.493	0.439	0.937
Observations	222	180	203		
48-59 months					
Score	10.881	10.396	10.428	0.090	0.109
	(0.180)	(0.217)	(0.213)		
Fail to pass pre-stage (%)	0.071	0.072	0.116	0.987	0.219
Altered (%)	0.143	0.230	0.152	0.070	0.832
Normal (%)	0.786	0.698	0.732	0.104	0.310
Observations	126	139	138		
60-72 months					
Score	11.024	11.284	10.522	0.421	0.211
	(0.230)	(0.223)	(0.341)		
Fail to pass pre-stage (%)	0.035	0.054	0.116	0.568	0.054
Altered (%)	0.129	0.068	0.145	0.198	0.782
Normal (%)	0.835	0.878	0.739	0.444	0.145
Observations	85	74	69		

Note: T-tests report comparisons between the control arm against NEP-B and NEP-I. Significance levels: *p<=10%, **p<=5%, ***p<=1%.

Table A4 describes parental beliefs, psychosocial well-being, and investments in children. The scale Ideas About Parenting, measuring parenting styles, does not show significant differences across treatment arms. We also do not find significant differences in parental perceived self-efficacy, or in perceived social support. Finally, we do not find significant differences in our measures of parental investments in children using a measure of home environments based on the Family Care Indicators (FCI), or the sub-scales of Nurturing and Discipline from the Parenting Behavior Checklist.

Table A4: Baseline balance: parental beliefs, mental health, and investments in children

Parental Indicators	(1) Control	(2) NEP-B	(3) NEP-I	t-test p-value	t-test p-value
Variable	Mean/SE	Mean/SE	Mean/SE	(1)-(2)	(1)-(3)
Authoritative style (IRT score)	-0.272	-0.287	-0.275	0.625	0.909
	(0.021)	(0.021)	(0.021)		
Authoritarian style (IRT score)	0.411	0.374	0.388	0.276	0.508
	(0.024)	(0.024)	(0.025)		
Permissive style (IRT score)	-0.538	-0.511	-0.539	0.189	0.963
	(0.014)	(0.015)	(0.014)		
Perceived Self-efficacy	64.220	64.173	64.545	0.911	0.444
	(0.302)	(0.298)	(0.299)		
Perceived Social Support	2.920	2.903	2.852	0.825	0.375
	(0.054)	(0.055)	(0.054)		
Parental Stress	29.943	30.361	29.676	0.427	0.610
	(0.373)	(0.370)	(0.368)		
Depression	40.222	41.072	39.600	0.136	0.271
	(0.406)	(0.399)	(0.394)		
Home Index (Family Care	0.810	0.771	0.791	0.168	0.504
Indicators scale)	(0.020)	(0.020)	(0.020)		
Socio-emotional stimulation (PBC	3.995	3.994	4.016	0.967	0.306
Nurturing Raw scale)	(0.015)	(0.014)	(0.014)		
Use of disciplinary strategies (PBC	2.729	2.733	2.692	0.877	0.180
Discipline raw scale)	(0.019)	(0.019)	(0.020)		
Observations	971	971	971		
F-test of joint significance (p-value)				0.599	0.341

Note: T-tests report comparisons between the control arm against NEP-B and NEP-I. Significance levels: *p<=10%, **p<=5%, ***p<=1%. F-test for the joint significance across all variables is reported at the bottom.

Figure A1 indicates that the outcomes of standardized language tests at baseline (PLS-IV) and of executive functions tests (DCCS) improve as the caregiver's educational attainment increases. The same is true if we plot the receptive language test TEVI-R using endline data. Figure A2 shows similar patterns for socio-emotional development: maladaptive behaviors (internalizing and externalizing behavioral problems) measured through the CBCL decrease as the caregiver's educational attainment increases, whereas adaptive behaviors, measured using the Battelle socio-personal scale, are positively related to the caregiver's educational attainment.

Figure A1: Baseline child cognitive development and primary caregiver education

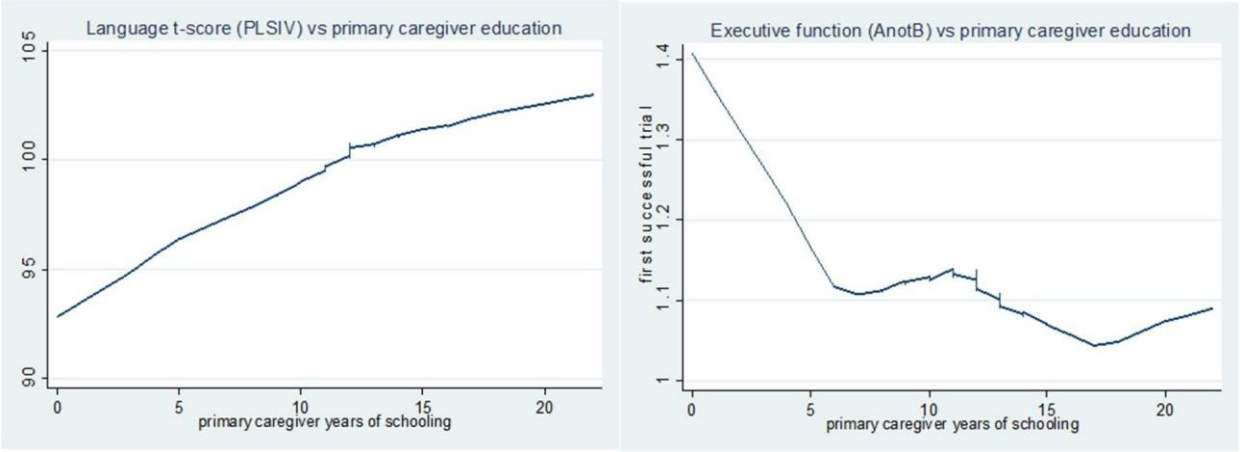


Figure A2: Baseline child behavior and primary caregiver education

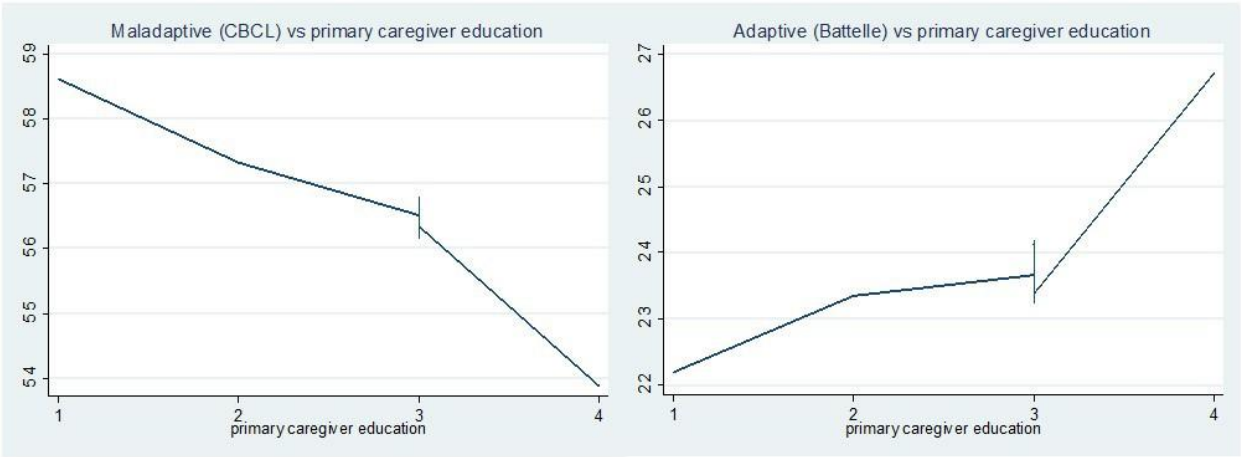


Figure A3 (right side) illustrates that positive cognitive stimulation practices measured through the HOME index are positively associated with the caregiver's educational attainment. Figure A3 (left side) shows that non-cognitive stimulation practices measured with the PBC nurturing scale also increase with the caregiver's education, while the use of harsh disciplinary practices decreases at higher levels of educational attainment.

Figure A3: Baseline parenting behaviors and primary caregiver education

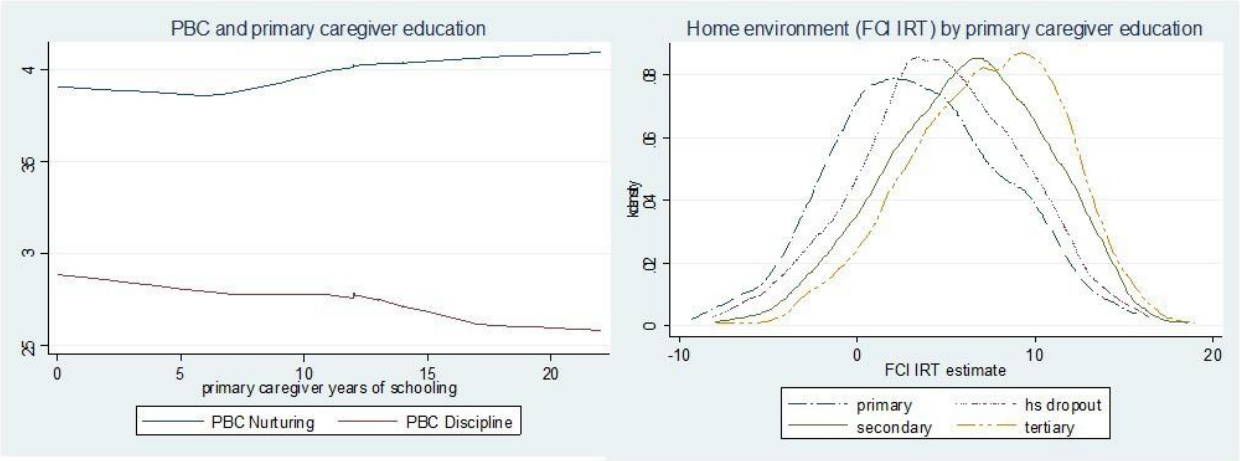
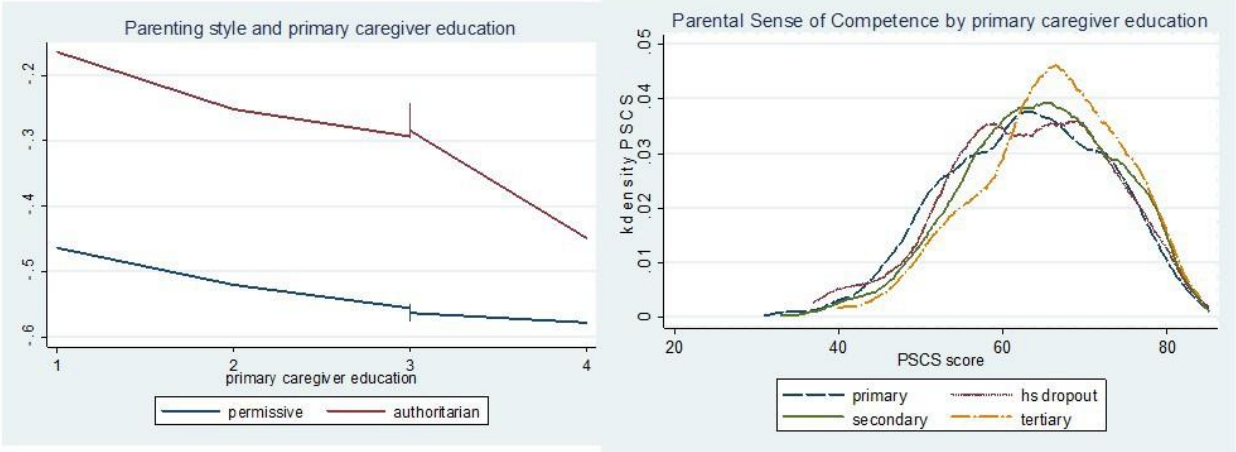


Figure A4 (left side) reveals important socio-economic gradients for parenting styles and parental beliefs. Authoritarian and permissive parental styles are more prevalent among parents with low educational attainment, in contrast with the authoritative style. Figure A4 (right side) illustrates that both perceived self-efficacy and social support increase as the caregiver's educational attainment increases.

Figure A4: Baseline parental beliefs



Appendix 2: Construction of the child and parental indices

Child Outcomes Indices

The construction of indices for child outcomes is motivated by two data issues. First, they allow correcting for measurement error in raw data potentially biasing the standard errors of key child outcome measures. Second, they are useful to combine different measures, some of them available for certain age groups but not others, into a single index, which can be estimated for all age groups as long as measures partially overlap across age groups. Table A5 shows the raw scores available for each scale and for each age group in our endline data, which justifies the construction of such indices.

Indices were obtained combining IRT and GSEM methods. However, we note that IRT and GSEM methods are equivalent, and both are based on Maximum Likelihood Estimation. The difference is that IRT methods are only applicable for discrete individual measures, while GSEM can be used with both continuous and discrete measures.

The Vocabulary Index at endline was estimated using the 116 items of the TEVI-R assessment using Item response theory (IRT) methods, which are better suited to predict latent variable scores using binary scale items in psychometric testing (Bock and Gibbons 2021, Lu, Thomas, and Zumbo 2005). The fundamental building block of IRT is the item characteristic curve (ICC), which links the latent ability, θ, to the probability a randomly drawn examinee of a given ability will answer the item correctly, P(θ). We estimated a Rasch two-parameter logistic (2PL) model, in which P(θ) varies with ability according to two parameters: a difficulty parameter measuring the item's overall difficulty, and a discrimination parameter, capturing how quickly the likelihood of success changes with respect to ability. Because most responses were incorrect in the last few items due to age characteristics of the sample, convergence was achieved including the 81 first items of the scale. If we use GSEM methods adapted to binary variables to estimate the vocabulary index, we obtain the same results.

For the Executive Function and Socioemotional indices, we relied on the raw scales (not items) to estimate latent constructs using Generalized Structural Equation Modelling (GSEM) methods in Stata. The advantage of this approach is that it allows us to construct indices for the full sample including all the information available even if some input measures have incomplete data due to age-eligibility criteria for the underlying scales. To describe the procedure, take for example the construction of the socioemotional index. From Table A5, we had data from 3 subscales of the Battelle Socio-personal test and 2 subscales of the CBCL test, but this data was incomplete: first, all these measures were collected for only one child in the household, and second, while the CBCL measures were collected for all age groups, the Battelle sub-scales were missing for older kids.

The full-sample index is then estimated for each age group separately including all the information available. For the first age group, where we have information for all the available scales but for

only one child per household (the target child), no restriction is applied besides standardizing the variance of the factor to 1. To predict the predicted latent factor for this group, GSEM treats missing data from socioemotional scales for children other than the target child assuming they are missing at random and multivariate normality. For the second age group, the model is estimated again with the available scales, but both the coefficient and the constant of the overlapping scales with the first age group (CBCL and Battelle Social Role) are fixed to be the values estimated for first age group, while the others are let free. For the third age group, only data from the CBCL is used. The Executive Function Index is obtained by estimating the GSEM model for all age groups simultaneously, as there is complete data.

As mentioned earlier, we also estimated a composite Child Development Index using all measures collected at endline. However, this single index was highly correlated with the Socioemotional Index so we left it out of our results. Similarly, we also estimate a single Baseline Child Development Index that includes all cognitive, language, and socioemotional measures at baseline (Battelle, CBCL, DCCS, and PLS-IV measuring receptive and expressive language).

Table A5: Raw child measures and indices of child outcomes.

		Vocabulary Index	Executive Function Index		Socio-emotional Index				
Age at Endline	Stat	TEVI-R	DCCS	Sustained Attention	Battelle - Peers Interaction	Battelle - Adults Interaction	Battelle - Social Role	CBCL-Extern.	CBCL – Intern.
3-4	mean	19.7	12.7	37.0	30.0	48.2	60.9	11.5	13.8
	N	*644*	*649*	*653*	*497*	*498*	*501*	*502*	*502*
4-5	mean	25.0	15.6	36.8	36.2	52.7	66.4	10.6	12.2
	N	*685*	*697*	*702*	*580*	*582*	*589*	*583*	*583*
5-6	mean	34.6	17.0	44.0	42.4	56.8	72.3	10.1	10.9
	N	*521*	*528*	*527*	*450*	*452*	*477*	*453*	*453*
6-7	mean	37.8	18.5	63.9			78.4	8.3	10.0
	N	*496*	*499*	*500*			*455*	*340*	*361*
7-8	mean	42.5	19.5	72.4			81.4	9.7	10.3
	N	*327*	*334*	*336*			*303*	*232*	*244*
>8	mean	45.0	19.5	84.3				10.7	10.9
	N	*173*	*172*	*175*				*123*	*124*
Total	mean	31.0	16.4	49.9	36.0	52.4	70.7	10.3	11.7
	N	*2846*	*2879*	*2893*	*1527*	*1532*	*2325*	*2233*	*2267*

Parental Outcomes Indices

We estimated individual indices of parenting behaviors using GSEM methods adapted to categorical individual items. For home stimulation (HOME inventory) we estimated a single latent factor as our exploratory factor analysis showed that only one factor from these items was relevant. For parental nurturing behaviors we also identified a single relevant nurturing factor from the Parental Behavioral Checklist sub-scale of Nurturing. Our exploratory factor analysis of the Parental Behavioral Checklist sub-scale of Discipline identified two relevant discipline factors, one for negative and another for positive discipline practices, so we adapted our GSEM estimation to estimate these two factors. All remaining individual parental measures are just the sum over the items composing each scale, as suggested by the creators.

In our main results, we present composite indices of parental behaviors, beliefs and well-being, which are all estimated again using GSEM methods for continuous variables. The behavioral index includes the estimated individual indices for home stimulation, nurturing and discipline behaviors. The beliefs index includes raw scores for parental perceived self-efficacy, perceived social support, perceived parental impact of own behavior on child development (PACOTIS scale), and the raw scores for three parenting styles (authoritarian, authoritative, permissive). The well-being index includes raw scores of parental stress and depression.

Standardizations

All indices for child outcomes are internally age-standardized in 2-month bands. All raw scales and indices obtained with these scales exhibit age gradients.

All the relevant measures and composite indices of parental investments in children, beliefs, and mental health were standardized relative to the full sample without considering age effects. Results are not sensitive to the adding or not age effects in the standardization.

Appendix 3: ITT and IV Impacts, sensitivity

Table A6: Sensitivity of ITT estimates to interviewer fixed effects and clustering

		(1) (age/gender) + Interviewer's fixed effects		(2) (age/gender) + Std. errors clustered at the household level	
Child Outcome	Obs.	NEP B	NEP I	NEP B	NEP I
Vocabulary Index	2895	0.070	0.099**††	0.062	0.090**††
		(0.046)	(0.046)	(0.045)	(0.045)
Executive Function Index	2895	-0.021	0.042	-0.024	0.034
		(0.045)	(0.045)	(0.045)	(0.045)
Socioemotional Index	2492	0.032	0.094*†	0.017	0.082*†
		(0.048)	(0.049)	(0.048)	(0.048)
Parental Outcome					
Behavioral Index	2545	0.063	0.108**††	0.062	0.098**††
		(0.046)	(0.046)	(0.047)	(0.047)
Beliefs Index	2545	0.037	0.111**††	0.049	0.124***††
		(0.048)	(0.047)	(0.048)	(0.048)
Well-being Index	2545	0.047	0.012	0.044	0.021
		(0.048)	(0.048)	(0.049)	(0.049)

Note: Each line within a column specification refers to a separate regression. All regressions control for child's age and sex, and for health center's fixed effects. In addition, Column 1 clusters standard errors at the household level and Column 2 controls for interviewer fixed-effects. Significance levels: *p<=10%, **p<=5% for individual hypotheses tests; ††p<=5% †p<=10% after accounting for multiple hypotheses tests.

Table A7: ITT estimates of individual measures of child outcomes

		(1) (age/gender)		(2) + caregiver's education, IQ and personality traits; hh'ld income, hh'ld size		(3) + baseline outcomes	
Child Individual Outcomes	Obs.	NEP-B	NEP-I	NEP-B	NEP-I	NEP-B	NEP-I
TEVI-R (language)	2895	0.070	0.099**††	0.070	0.099**††	0.070	0.099**††
		(0.044)	(0.045)	(0.044)	(0.045)	(0.044)	(0.045)
DCCS (cognitive flexibility)	2890	-0.036	-0.070	-0.036	-0.070	-0.036	-0.070
		(0.044)	(0.044)	(0.044)	(0.044)	(0.044)	(0.044)
Leiter-R (sustained attention)	2893	-0.022	0.014	-0.022	0.014	-0.022	0.014
		(0.043)	(0.044)	(0.043)	(0.044)	(0.043)	(0.044)
CBCL total (behavior)	1840	-0.023	-0.049	-0.023	-0.049	-0.023	-0.049

		(0.051)	(0.050)	(0.051)	(0.050)	(0.051)	(0.050)
CBCL externalization	1971	-0.023	-0.020	-0.023	-0.020	-0.023	-0.020
		(0.050)	(0.050)	(0.050)	(0.050)	(0.050)	(0.050)
CBCL internalization	1887	-0.031	-0.020	-0.031	-0.020	-0.031	-0.020
		(0.049)	(0.049)	(0.049)	(0.049)	(0.049)	(0.049)
CBCL Emotional	1991	0.009	-0.028	0.009	-0.028	0.009	-0.028
		(0.052)	(0.052)	(0.052)	(0.052)	(0.052)	(0.052)
CBCL Anxious	2008	0.013	0.014	0.013	0.014	0.013	0.014
		(0.051)	(0.051)	(0.051)	(0.051)	(0.051)	(0.051)
CBCL Somatic	1958	-0.017	-0.007	-0.017	-0.007	-0.017	-0.007
		(0.052)	(0.052)	(0.052)	(0.052)	(0.052)	(0.052)
CBCL Withdrawn	2119	0.001	0.023	0.001	0.023	0.001	0.023
		(0.051)	(0.050)	(0.051)	(0.050)	(0.051)	(0.050)
CBCL Sleep problems	1947	0.108**	0.047	0.108**	0.047	0.108**	0.047
		(0.055)	(0.054)	(0.055)	(0.054)	(0.055)	(0.054)
CBCL Attention problems	2022	-0.035	-0.031	-0.035	-0.031	-0.035	-0.031
		(0.053)	(0.052)	(0.053)	(0.052)	(0.053)	(0.052)
CBCL Aggressive	2011	0.031	0.002	0.031	0.002	0.031	0.002
		(0.053)	(0.052)	(0.053)	(0.052)	(0.053)	(0.052)
Battelle Social Role	2325	0.057	0.096**††	0.057	0.096**††	0.057	0.096**††
		(0.048)	(0.048)	(0.048)	(0.048)	(0.048)	(0.048)
Battelle Interaction w/Adults	1532	0.072	0.152**††	0.072	0.152**††	0.072	0.152**††
		(0.062)	(0.062)	(0.062)	(0.062)	(0.062)	(0.062)
Battelle Interaction w/Peers	1521	0.002	0.083	0.002	0.083	0.002	0.083
		(0.063)	(0.064)	(0.063)	(0.064)	(0.063)	(0.064)

Note: Each line within a column specification refers to a separate regression. All regressions control for child's age and sex, and for health center's fixed effects. All scales are constructed adding raw item responses except for the TEVI-R score which is estimated with IRT models. Significance levels: *p<=10%, **p<=5% for individual hypotheses tests; ††p<=5% †p<=10% after accounting for multiple hypotheses tests.

Table A8: ITT estimates of individual measures of parental outcomes

	(1) (age/gender)		(2) + caregiver's education, IQ and personality traits; hh'ld income, hh'ld size	
Parental Practices	NEP B	NEP I	NEP B	NEP I
Home Index	0.064	0.085*†	0.082*	0.078*†
	(0.046)	(0.046)	(0.044)	(0.044)
PBC Nurturing Index	0.037	0.077*†	0.044	0.071
	(0.046)	(0.046)	(0.046)	(0.046)
PBC Negative Discipline Index	-0.020	-0.046	-0.006	-0.035
	(0.047)	(0.047)	(0.043)	(0.043)
PBC Positive Discipline Index	0.070	0.101**††	0.056	0.083*†

	(0.047)	(0.048)	(0.044)	(0.044)
Parental Beliefs				
Perceived Self-efficacy	0.033	0.101**††	0.023	0.086*†††*
	(0.048)	(0.048)	(0.042)	(0.042)
PACOTIS scale	0.069	0.105**††	0.066	0.092**††
	(0.047)	(0.047)	(0.045)	(0.045)
Perceived Family Support	-0.079	0.003	-0.064	0.013
	(0.049)	(0.049)	(0.048)	(0.048)
Perceived Friends Support	0.070	0.083*	0.076*	0.071
	(0.047)	(0.047)	(0.045)	(0.045)
Perceived Support from Others	-0.013	0.014	-0.006	0.012
	(0.047)	(0.047)	(0.047)	(0.047)
Authoritative Style	0.027	0.029	0.030	0.036
	(0.047)	(0.048)	(0.046)	(0.046)
Authoritarian Style	-0.002	0.061	-0.016	0.055
	(0.048)	(0.048)	(0.047)	(0.047)
Permissive Style	-0.042	0.004	-0.039	0.008
	(0.047)	(0.048)	(0.047)	(0.047)
Parental Mental Health				
Parental Stress	0.044	-0.011	0.033	-0.015
	(0.048)	(0.048)	(0.045)	(0.045)
Depression	0.041	0.033	0.037	0.034
	(0.048)	(0.048)	(0.043)	(0.043)
Observations	2545			

Note: Each line within a column specification refers to a separate regression. All regressions control for child's age and sex, and for health center's fixed effects. PACOTIS scales measures parental beliefs about the importance of own behaviors for child development. Home, Nurturing and Discipline indices obtained from raw data using Principal Component Analysis (PCA). Significance levels: *p<=10%, **p<=5% for individual hypotheses tests; ††p<=5% †p<=10% after accounting for multiple hypotheses tests.

Table A9: First stage (extended table with all controls)

	Participation NEP-B	Participation NEP-I
	Coef (SE)	Coef (SE)
NEP-B	0.202***	
	(0.012)	
NEP-I		0.263***
		(0.013)
Child's age at baseline (base: 0-12 mo.)		
13-24 mo.	0.005	0.029
	(0.018)	(0.019)
25-36 mo.	0.013	0.045**
	(0.020)	(0.020)
37-48 mo.	-0.006	0.014
	(0.020)	(0.021)

49-60 mo.	-0.017	0.024
	(0.024)	(0.025)
Girls	0.010	0.017
	(0.012)	(0.013)
HH Incomes (base: q1)		
q2	-0.007	0.039*
	(0.019)	(0.020)
q3	-0.005	0.001
	(0.020)	(0.020)
q4	-0.021	-0.012
	(0.020)	(0.021)
q5	-0.012	0.011
	(0.022)	(0.023)
Caregiver Education (base: Primary)		
High School Dropout	-0.009	-0.006
	(0.020)	(0.021)
High School Degree	0.031*	-0.010
	(0.016)	(0.017)
College	0.057**	0.036
	(0.023)	(0.024)
Number of HH members	-0.004	0.001
	(0.004)	(0.005)
Single mother	-0.033***	--0.031**
	(0.013)	(0.013)
Number of younger siblings	0.015	-0.010
	(0.014)	(0.014)
Caregiver active at baseline	--0.030**	-0.025*
	(0.014)	(0.015)
Caregiver works full-time at baseline	0.015	0.005
	(0.019)	(0.019)
Observations	2528	2528

Note: All regressions control for health center's fixed effects. Significance levels: *p<10%, **p<5%, ***p<1%

Table A10: IV estimates of child outcomes, with added controls

		(1) (age/gender) + Std. errors clustered at the household level		(2) (age/gender) + Interviewer's fixed effects	
Child Outcome	Obs.	NEP B	NEP I	NEP B	NEP I
Vocabulary Index	2895	0.386	0.425**††	0.343	0.380**††
		(0.251)	(0.196)	(0.244)	(0.188)
Executive Function Index	2895	-0.107	0.149	-0.127	0.119
		(0.244)	(0.189)	(0.243)	(0.187)
Socioemotional Index	2492	0.181	0.381*†	0.101	0.330*†
		(0.257)	(0.209)	(0.260)	(0.200)
Parental Outcome					

Behavioral Index	2545	0.371	0.463**††	0.361	0.415**††
		(0.256)	(0.199)	(0.258)	(0.198)
Beliefs Index	2545	0.220	0.465**††	0.286	0.517**††
		(0.264)	(0.205)	(0.265)	(0.203)
Well-being Index	2545	0.256	0.080	0.236	0.109
		(0.264)	(0.204)	(0.267)	(0.205)

Note: Each line reports estimates from separate 2SLS regressions using randomization status as instrumental variables for attending at least one session. Dependent variables are indices of child and parental outcomes measured at follow-up. Effect sizes use internal age-standardization to the full sample (mean 0, SD=1). All regressions control for child sex and age, and for health center's fixed effects. In addition, Column 1 clusters standard errors at the household level and Column 2 controls for interviewer fixed-effects. Significance levels: *p<=10%, **p<=5% for individual hypotheses tests; ††p<=5% †p<=10% after accounting for multiple hypotheses tests.

Table A11: IV estimates of child outcomes using number of sessions attended

		(1) (age/gender)		(2) + caregiver's education, IQ and personality traits; hh'ld income, hh'ld size		(3) + baseline outcomes	
Child Outcome	Obs.	NEP B	NEP I	NEP B	NEP I	NEP B	NEP I
Vocabulary Index	2895	0.077	0.056**††	0.083*	0.059**††	0.082*	0.059**††
		(0.048)	(0.025)	(0.048)	(0.025)	(0.048)	(0.025)
Exec. Function Index	2895	-0.020	0.018	-0.023	0.016	-0.024	0.016
		(0.047)	(0.025)	(0.047)	(0.025)	(0.046)	(0.024)
Socioemotional Index	2492	0.035	0.048*†	0.044	0.049*†	0.045	0.050**††
		(0.049)	(0.026)	(0.048)	(0.026)	(0.048)	(0.025)

Note: Each line reports estimates from separate 2SLS regressions using randomization status as instrumental variables for the number of sessions attended in each treatment arm. Dependent variables are indices of child development outcomes measured at follow-up. Effect sizes use internal age-standardization to the full sample (mean 0, SD=1). All regressions control for child sex and age, and for health center's fixed effects. Significance levels: *p<=10%, **p<=5% for individual hypotheses tests; ††p<=5% †p<=10% after accounting for multiple hypotheses tests.

Table A12: IV estimates of parental outcomes using number of sessions attended

		(1) (age/gender)		(2) + caregiver's education, IQ and personality traits; hh'ld income, hh'ld size	
Parental Outcome	Obs.	NEP B	NEP I	NEP B	NEP I
Behavioral Index	2545	0.068	0.060**††	0.072	0.055**††
		(0.050)	(0.027)	(0.047)	(0.025)
Beliefs Index	2545	0.041	0.059**††	0.031	0.048**††
		(0.051)	(0.027)	(0.044)	(0.024)
Well-being Index	2545	0.050	0.012	0.044	0.012

		(0.051)	(0.027)	(0.045)	(0.024)

Note: Each line reports estimates from separate 2SLS regressions using randomization status as instrumental variables for the number of sessions attended in each treatment arm. Dependent variables are indices of parental outcomes measured at follow-up. Effect sizes use internal age-standardization to the full sample (mean 0, SD=1). All regressions control for child sex and age, and for health center's fixed effects. Significance levels: *p<=10%, **p<=5% for individual hypotheses tests; ††p<=5% †p<=10% after accounting for multiple hypotheses tests.

Table A13: IV es estimates of individual measures of child outcomes

		(1) (age/gender)		(2) + caregiver's education, IQ and personality traits; hh'ld income and size		(3) + baseline outcomes	
Child Individual Outcomes	Obs.	NEP-B	NEP-I	NEP-B	NEP-I	NEP-B	NEP-I
TEVI-R (language)	2895	0.386	0.425**	0.417*	0.445**	0.415*	0.447**
		(0.243)	(0.189)	(0.242)	(0.189)	(0.241)	(0.189)
DCCS (cognitive flexibility)	2890	-0.199	-0.290	-0.198	-0.303	-0.199	-0.282
		(0.240)	(0.187)	(0.239)	(0.188)	(0.228)	(0.179)
Leiter-R (sustained attention)	2893	-0.117	0.042	-0.136	0.031	-0.142	0.035
		(0.236)	(0.184)	(0.235)	(0.184)	(0.234)	(0.183)
CBCL total (behavior)	1840	0.292	0.398**	0.314	0.403**	0.326	0.417**
		(0.248)	(0.201)	(0.245)	(0.200)	(0.244)	(0.199)
CBCL externalization	1971	0.354	0.616**	0.328	0.523**	0.312	0.524**
		(0.312)	(0.262)	(0.301)	(0.253)	(0.300)	(0.253)
CBCL internalization	1887	0.008	0.317	0.020	0.283	-0.005	0.287
		(0.314)	(0.267)	(0.304)	(0.258)	(0.302)	(0.257)
CBCL Emotional	1991	-0.132	-0.201	-0.168	-0.257	-0.173	-0.260
		(0.280)	(0.210)	(0.266)	(0.200)	(0.266)	(0.200)
CBCL Anxious	2008	-0.128	-0.092	-0.209	-0.161	-0.212	-0.163
		(0.274)	(0.210)	(0.263)	(0.203)	(0.262)	(0.203)
CBCL Somatic	1958	-0.171	-0.096	-0.242	-0.143	-0.240	-0.143
		(0.273)	(0.203)	(0.261)	(0.196)	(0.261)	(0.195)
CBCL Withdrawn	2119	0.047	-0.102	-0.055	-0.168	-0.056	-0.170
		(0.292)	(0.218)	(0.280)	(0.211)	(0.280)	(0.211)
CBCL Sleep problems	1947	0.076	0.060	0.004	0.029	-0.001	0.024
		(0.286)	(0.209)	(0.276)	(0.203)	(0.275)	(0.203)
CBCL Attention problems	2022	-0.088	-0.038	-0.152	-0.074	-0.151	-0.074
		(0.278)	(0.216)	(0.273)	(0.213)	(0.273)	(0.213)
CBCL Aggressive	2011	0.006	0.088	-0.035	0.086	-0.033	0.083
		(0.275)	(0.209)	(0.267)	(0.204)	(0.267)	(0.204)
Battelle Social Role	2325	0.577*	0.233	0.560*	0.210	0.558*	0.207
		(0.297)	(0.220)	(0.292)	(0.218)	(0.292)	(0.218)
Battelle Interaction with Adults	1532	-0.190	-0.142	-0.270	-0.185	-0.276	-0.191
		(0.285)	(0.220)	(0.280)	(0.217)	(0.279)	(0.216)

Battelle Interaction with Peers	1521	0.165	0.026	0.090	-0.037	0.087	-0.038
		(0.286)	(0.219)	(0.277)	(0.214)	(0.277)	(0.214)

Note: Each line reports estimates from separate 2SLS regressions using randomization status as instrumental variables for attending a least one session. Dependent variables measured at follow-up are scores constructed adding raw item responses except for the TEVI-I score which is estimated with IRT models. Effect sizes use internal age-standardization to the full sample (mean 0, SD=1). All regression control for child sex and age, and for health center's fixed effects. Significance levels: *p<=10%, **p<=5% for individual hypothese tests; ††p<=5% †p<=10% after accounting for multiple hypotheses tests.

Table A14: IV es estimates of individual measures of parental outcomes

	(1) (age/gender)		(2) + caregiver's education, IQ and personality traits; hh'ld income, hh'ld size	
Parental Practices	NEP B	NEP I	NEP B	NEP I
Home Index	0.373	0.372*	0.455*	0.361*
	(0.257)	(0.200)	(0.244)	(0.191)
PBC Nurturing Index	0.230	0.325*	0.247	0.305
	(0.255)	(0.196)	(0.251)	(0.196)
PBC Negative Discipline Index	-0.105	-0.196	-0.036	-0.140
	(0.257)	(0.200)	(0.233)	(0.182)
PBC Positive Discipline Index	0.394	0.440**	0.310	0.357*
	(0.265)	(0.206)	(0.243)	(0.190)
Parental Beliefs				
Perceived Self-efficacy	0.196	0.417**	0.132	0.346*
	(0.261)	(0.204)	(0.231)	(0.181)
PACOTIS scale	0.387	0.455**	0.367	0.401**
	(0.259)	(0.201)	(0.246)	(0.192)
Perceived Family Support	-0.420	-0.040	-0.345	0.008
	(0.270)	(0.210)	(0.262)	(0.204)
Perceived Friends Support	0.386	0.366*	0.418*	0.327*
	(0.263)	(0.205)	(0.250)	(0.195)
Perceived Support from Others	-0.075	0.043	-0.033	0.041
	(0.260)	(0.203)	(0.254)	(0.198)
Authoritative Style	0.151	0.132	0.164	0.157
	(0.261)	(0.203)	(0.249)	(0.195)
Authoritarian Style	0.001	0.243	-0.082	0.203
	(0.261)	(0.203)	(0.255)	(0.200)
Permissive Style	-0.226	-0.013	-0.212	0.004
	(0.261)	(0.203)	(0.258)	(0.202)
Parental Mental Health				
Parental Stress	0.232	-0.015	0.176	-0.036
	(0.263)	(0.204)	(0.245)	(0.191)
Depression	0.226	0.158	0.203	0.158
	(0.264)	(0.205)	(0.233)	(0.182)

Observations	2545	

Note: Each line reports estimates from separate 2SLS regressions using randomization status as instrumental variables for the number of sessions attended in each treatment arm. Dependent variables are indices of parental outcomes measured at follow-up. Effect sizes use internal age-standardization to the full sample (mean 0, SD=1). All regressions control for child sex and age, and for health center's fixed effects. Significance levels: *p<=10%, **p<=5% for individual hypotheses tests; ††p<=5% †p<=10% after accounting for multiple hypotheses tests.

Table A15: Comparison between OLS and IV estimates of main outcomes

		Participation NEP-B		Participation NEP-I	
	Obs.	OLS	IV	OLS	IV
Vocabulary Index	2895	0.160***	0.386	0.120**	0.425**
		(0.061)	(0.243)	(0.057)	(0.189)
Socioemotional Index	2492	0.036	0.181	0.127**	0.381*
		(0.065)	(0.257)	(0.061)	(0.201)
Parental Behavioral Index	2545	0.094	0.353	0.083	0.460**
		(0.063)	(0.258)	(0.060)	(0.201)
Parental Beliefs Index	2545	0.043	0.211	0.089	0.457**
		(0.065)	(0.263)	(0.061)	(0.205)

Note: Each line reports estimates from separate OLS or 2SLS regressions. Dependent variables are indices of key child and parental outcomes measured at follow-up. Effect sizes use internal age-standardization to the full sample (mean 0, SD=1). All regressions control for child sex and age, and for health center's fixed effects. Significance levels: *p<=10%, **p<=5% testing individual hypotheses.

Table A16: Selection on Unobservables

	(1) Participants NEP-B	(2) Participants NEP-I	(3) Non-Participants
Panel A: Parental Behavioral Index			
NEP-B	0.119	-0.166	0.022
	(0.215)	(0.240)	(0.053)
NEP-I	0.188	0.081	0.081
	(0.299)	(0.184)	(0.054)
p-value joint test NEP-B=NEP-I=0	0.787	0.438	0.315
Observations	298	351	1896
Panel B: Parental Beliefs Index			
NEP-B	0.089	-0.250	0.009
	(0.200)	(0.258)	(0.055)
NEP-I	0.218	-0.062	0.104*
	(0.279)	(0.198)	(0.056)
p-value joint test NEP-B=NEP-I=0	0.734	0.590	0.135
Observations	298	351	1896

Note: Each column in each panel (a, b or c) represents a separate OLS regression. Column 1 uses the sample of participants into the NEP-B program only. Column 2 uses the sample of participants into the NEP-I only. Column 3 uses the sample of non-participants only. Dependent variables are indices of key child and parental outcomes measured at follow-up. Effect sizes use internal age-standardization to the full sample (mean 0, SD=1). All regressions control for

child sex and age, and for health center's fixed effects. Significance levels: *p<=10%, **p<=5% testing individual hypotheses.

Appendix 4: Attrition

In this section, we examine the potential importance of selective attrition between baseline and follow-up. Table A17 shows that there is some degree of selective attrition when comparing NEP-B against the control group (Column 1). In the next columns of Table A17, we investigate whether some key outcomes of the study as well as SES variables can explain differential attrition by treatment arm. First, we find evidence of a positive and significant interaction between receptive language at baseline and NEP-B, which is fully explained by a higher language score among non-attrites vs. attrites within the Control Group. And second, there is a negative and significant interaction between household income at baseline and NEP-I. Interactions between outcomes and NEP-I are never significant.

We correct for the potential bias arising from selective attrition using three approaches. First, we estimate "upper" and "lower" Lee Bounds to correct for attrition bias (Lee 2009), which contrary to parametric approaches requires only the assumptions of random assignment of treatment and monotonicity on how treatment status affects selection, and which rests on a trimming procedure from above and below. Table A18 shows the estimated bounds for each of the four child outcomes. Across all outcomes, the "upper" bounds are statistically larger than "lower" bounds. Among the three child outcomes with statistically significant impacts in Table 3, the "lower" bounds for the Child Development and Vocabulary Indices are above zero, and for the Socioemotional Index is negative but close to zero.

Second, we further test for potential biases due to selective attrition in child outcomes using Inverse Probability Weighting (IPW) techniques in a two-stage estimation approach (Robins, Rotnitzky, and Zhao 1994). In the first stage, we use logistic models to estimate the probability that a child is a non-attrite at endline as a function of baseline variables including child sex, age and birth order, caregiver's age, gender, household income and composition, baseline language and executive function and treatment assignment. The inverse of the predicted probabilities obtained from the first stage is used as weights in the second stage outcome regression so that a larger weight is given to participants who are underrepresented in the sample as a result of attrition. Observations with implausible large weights (over 50) were dropped from the sample in the second stage regressions (5% of the data). Table A19 shows that the IPW-adjusted regressions with age-standardized child outcomes are statistically equivalent to our main findings from Table 3.

Our third approach follows the method proposed by (Angrist, Bettinger, and Kremer 2006) who estimate Tobit regressions for a censored outcome for different percentiles of the distribution of the latent variable, assigning the value of the outcome at the percentile for missing values, and the observed outcome for values above the percentile. The idea is to test for the stability of the coefficient of interest across regressions when the percentile is increased. Tables A20 and A21 present the Tobit regression outcomes censoring the outcome variable at different percentiles and show that the impacts would remain robust regardless of the percentile chosen to censor the data.

Table A17: Attrition: interaction with treatment arms and baseline variables

Dep. Var: Attrition	Treatment	Treatment x[baseline variable]					
		Language	Ex. Function	Socio-emotional	Behavior	Education	Income
NEP-B	-0.047***	-0.265***	-0.074***	-0.001	-0.027	-0.052	-0.007
	(0.016)	(0.098)	(0.028)	(0.036)	(0.041)	(0.044)	(0.037)
NEP-I	-0.027*	-0.165*	-0.071**	-0.018	-0.026	-0.062	0.040
	(0.016)	(0.098)	(0.028)	(0.035)	(0.040)	(0.045)	(0.037)
[baseline variable]		-0.002***	-0.008***	-0.001	-0.001	0.001	0.001
		(0.001)	(0.003)	(0.001)	(0.001)	(0.012)	(0.008)
NEP-B x [baseline variable]		0.002**	0.004	-0.002	-0.000	0.002	-0.014
		(0.001)	(0.004)	(0.001)	(0.001)	(0.016)	(0.011)
NEP-I x [baseline variable]		0.001	0.007*	-0.000	0.000	0.010	-0.023**
		(0.001)	(0.004)	(0.001)	(0.001)	(0.016)	(0.011)
Observations	3579	3188	3568	3509	2282	3562	3579

Note: Each column represents a separate regression where a baseline variable is interacted with treatment dummies., controlling for health center's fixed effects. Standard errors are reported in parentheses. Significance levels: *p<10%, **p<5%.

Table A18: Lee Bounds (z-index)

Bound	Vocabulary Index		Exec. Function Index		Socioemotional Index	
	NEP_B	NEP_I	NEP_B	NEP_I	NEP_B	NEP_I
lower	-0.043	0.043	-0.122**	-0.005	-0.121*	-0.022
	(0.057)	(0.060)	(0.058)	(0.061)	(0.063)	(0.068)
upper	0.239***	0.207***	0.140**	0.157**	0.163**	0.165**
	(0.061)	(0.067)	(0.061)	(0.070)	(0.066)	(0.069)

Note: Each column represents the estimation of Lee Bounds for a given child outcome and a treatment arm compared with the Control group. All specifications control for child's age and sex, baseline language and executive function outcomes, household income, family composition, and for health center's fixed effects. Significance levels: *p<=10%, **p<=5% and ***p<=5% testing individual hypotheses

Table A19: Inverse Probability Weighting (z-index)

Child Outcome	Observations	NEP B	NEP I	P value for B=I
Vocabulary Index	2467	0.080**	0.073*	0.838
		(0.039)	(0.038)	
Executive Function Index	2467	-0.001	0.031	0.273
		(0.028)	(0.028)	
Socioemotional Index	2171	0.046	0.072**	0.435
		(0.035)	(0.035)	

Note: Each line within a column specification refers to a separate weighted regression. Weights were obtained as the inverse of the predicted probability that a child was a non-attrite, which was estimated in a first stage as a function of

baseline variables including child sex, age and birth order, caregiver's age, baseline language and executive function scores, household income and composition, and treatment assignment. All second-stage regressions control for child's age and sex, and for health center's fixed effects. Significance levels: *p<=10%, **p<=5% testing individual hypotheses.

Table A20: Tobit regression for the impacts on child vocabulary index at endline

	OLS	TOBIT censored at:					
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
		5%	10%	20%	30%	40%	50%
NEP-B	0.076*	0.138***	0.166***	0.076**	0.109***	0.094**	0.072*
	(0.045)	(0.042)	(0.049)	(0.030)	(0.042)	(0.041)	(0.042)
NEP-I	0.099**	0.114***	0.139***	0.068**	0.096**	0.082**	0.079*
	(0.045)	(0.042)	(0.049)	(0.030)	(0.042)	(0.041)	(0.042)
Observations	2895						

Note: Each column refers to a separate regression that controls for child sex and gender, and for health center's fixed effects. We adopt the procedure by Angrist, Bettinger, and Kremer (2006), whereby the sample of children with observed vocabulary index at endline is censored. Column (1) reports the main impact of NEP without adjusting for censoring. Columns (2)-(7) assume that the data is censored at the 5th, 10th, 20th, 30th, 40th and 50th percentile and is estimated with a Tobit model. Significance levels: *p<10%, **p<5%, ***p<1%.

Table A21: Tobit regression for the impacts on child socioemotional index at endline

	OLS	TOBIT censored at:					
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
		5%	10%	20%	30%	40%	50%
NEP-B	0.032	0.086*	0.065	0.035	0.066	0.008	0.016
	(0.047)	(0.046)	(0.041)	(0.034)	(0.050)	(0.024)	(0.048)
NEP-I	0.094**	0.110**	0.091**	0.063*	0.088*	0.042*	0.071
	(0.047)	(0.045)	(0.041)	(0.033)	(0.050)	(0.024)	(0.048)
Observations	2492						

Note: Each column refers to a separate regression that controls for child sex and gender, and for health center's fixed effects. We adopt the procedure by Angrist, Bettinger, and Kremer (2006), whereby the sample of children with observed socioemotional index at endline is censored. Column (1) reports the main impact of NEP without adjusting for censoring. Columns (2)-(7) assume that the data is censored at the 5th, 10th, 20th, 30th, 40th and 50th percentile and is estimated with a Tobit model. Significance levels: *p<10%, **p<5%, ***p<1%.

Appendix 5: Heterogeneous Treatment Effects

We examined heterogeneous impacts of NEP on child outcomes and quality of the home environment along several dimensions of caregiver's outcomes including education, cognitive ability, depressive symptoms, and personality traits, as well as by child sex, age, and outcomes at baseline.

Heterogeneity by Caregiver Characteristics/Outcomes

Treatment effects in child outcomes among children of lower educated caregivers (high school dropouts or less) are generally higher than those with more education (Table A22), although we there are significant differences only in NEP-I for the Socioemotional index. Differences in the Executive Function Index are large and close to a rejection of equality of impacts across subgroups. This general pattern is also observed for NEP-B, but differences across education groups are never significant.

Table A22: Heterogeneity of impact: caregiver education

		NEP-B			NEP-I		
		(1)	(2)		(3)	(4)	
	Obs.	Low Education	High Education	p-test (1)=(2)	Low Education	High Education	p-test (3)=(4)
Vocabulary Index	2895	0.151**	0.035	0.216	0.125	0.082	0.652
		(0.073)	(0.057)		(0.076)	(0.056)	
Exec. Function Index	2895	0.056	-0.063	0.202	0.134*	-0.017	0.110
		(0.072)	(0.057)		(0.075)	(0.056)	
Socioemotional Index	2492	0.119	-0.008	0.203	0.202**	0.021	0.072
		(0.078)	(0.060)		(0.080)	(0.059)	

Note: Low education group: caregivers with less than a high school degree education; High education group: caregivers with a high school degree or more. Each row reports estimates from a separate regression that includes interaction terms of intervention assignment with education category (low or high). Outcome variables are age-standardized. Significance levels: *p<10%, **p<5%, ***p<1%.

Impacts in child outcomes by caregiver IQ are measured using the Wechsler Adult Intelligence Scale (WAIS) (Table A23). Caregivers with Low IQ are those below the median of cognitive ability and those with high IQ those above the median. Impacts of NEP-B are significantly larger among the most disadvantaged group for the Vocabulary and Socioemotional indices, but not for the Executive Function Index. In NEP-I, impacts are also larger among caregivers with lower IQ but are statistically significant only for the Vocabulary Index.

Table A23: Heterogeneity of impact: caregiver cognition (IQ)

		NEP-B			NEP-I		
		(1)	(2)	p-test	(3)	(4)	p-test
	Obs.	Low IQ	High IQ	(1)=(2)	Low IQ	High IQ	(3)=(4)
Vocabulary Index	2895	0.161**	-0.000	0.076	0.197***	0.011	0.042
		(0.063)	(0.064)		(0.064)	(0.064)	
Exec. Function Index	2895	-0.057	0.020	0.391	0.006	0.081	0.411

		(0.063)	(0.063)		(0.063)	(0.064)	
Socioemotional Index	2492	0.129*	-0.066	0.045	0.145**	0.039	0.278
		(0.068)	(0.068)		(0.068)	(0.068)	

Note: Caregiver cognition (IQ) is measured with the Wechsler Adult Intelligence Scale (WAIS). Low IQ is defined as below the median IQ in the NEP sample, and High IQ is above the median. Each row reports estimates from a separate regression that includes interaction terms of intervention assignment with cognition category (low or high). Outcome variables are age-standardized. Significance levels: *p<10%, **p<5%, ***p<1%.

Treatment effects by caregiver's depressive symptoms were measured using the CES-D scale (Table A24). Using CES-D standards, caregivers with a score of 16 or more are considered to have depressive symptoms and are considered "Depressed", and those with scores lower than 16 are "Not Depressed". While results for the Executive Function and Socioemotional indices are generally larger among Not Depressed caregivers and the opposite is true for the Vocabulary Index, none of these differences are statistically significant.

Table A24: Heterogeneity of impact: caregiver's depressive symptoms

		NEP-B			NEP-I		
	Obs.	(1) Not Depressed	(2) Depressed	p-test (1)=(2)	(3) Not Depressed	(4) Depressed	p-test (3)=(4)
Vocabulary Index	2895	0.073	0.087	0.888	0.072	0.171**	0.338
		(0.053)	(0.085)		(0.053)	(0.087)	
Exec. Function Index	2895	0.002	-0.077	0.428	0.061	-0.019	0.433
		(0.053)	(0.084)		(0.052)	(0.086)	
Socioemotional Index	2492	0.055	-0.006	0.571	0.125**	0.008	0.283
		(0.055)	(0.091)		(0.055)	(0.093)	

Note: Caregiver depressive symptoms is measured with the Center for Epidemiologic Studies Depression Scale (CES-D). Depressed caregivers are those with a score of 16 or more, Each row reports estimates from a separate regression that includes interaction terms of intervention assignment with the depression dummy. Outcome variables are age-standardized. Significance levels: *p<10%, **p<5%, ***p<1%.

Heterogeneity by Child Characteristics/Outcomes

Impacts in child outcomes by child gender are never statistically significant (Table A25), though the effects in Vocabulary are generally larger for girls than boys in both treatment arms, and larger for boys in the other outcomes only in NEP-I.

Table A25: Heterogeneity of impact: child's sex

		NEP-B			NEP-I		
	Obs.	(1) Boys	(2) Girls	p-test (1)=(2)	(3) Boys	(4) Girls	p-test (3)=(4)
Vocabulary Index	2895	0.019	0.140**	0.183	0.076	0.128*	0.577
		(0.062)	(0.065)		(0.062)	(0.067)	
Exec. Function Index	2895	-0.026	-0.015	0.901	-0.005	0.096	0.269
		(0.061)	(0.064)		(0.062)	(0.066)	
Socioemotional Index	2492	0.021	0.044	0.811	0.140**	0.039	0.303
		(0.066)	(0.069)		(0.066)	(0.070)	

Note: Each row reports estimates from a separate regression that includes interaction terms of intervention assignment with a dummy variable for child's sex. Outcome variables are age-standardized. Significance levels: *p<10%, **p<5%, ***p<1%.

We examine impacts on child outcomes by children's age splitting the sample between children younger and older than 3 years old at baseline (Table A26). Generally, impacts across child outcomes, with the exception of vocabulary, are larger among younger children in both treatment arms, but differences are never statistically significant. Changing the definition of younger children to those who were 2 years old or less at baseline delivered the same results (not shown).

Table A26: Heterogeneity of impact: child's age

		NEP-B			NEP-I		
Treatment effects by child's age	Obs.	(1) Younger	(2) Older	p-test (1)=(2)	(3) Younger	(4) Older	p-test (3)=(4)
Vocabulary Index	2895	0.064	0.095	0.742	0.063	0.171**	0.264
		(0.055)	(0.079)		(0.055)	(0.078)	
Exec. Function Index	2895	0.025	-0.114	0.145	0.075	-0.030	0.271
		(0.054)	(0.078)		(0.055)	(0.077)	
Socioemotional Index	2492	0.073	-0.044	0.255	0.115*	0.051	0.533
		(0.059)	(0.082)		(0.060)	(0.081)	

Note: Younger children are those of 36 months of age or less at baseline (66.7% of the sample). Each row reports estimates from a separate regression that includes interaction terms of intervention assignment with a dummy variable for younger/older children. Outcome variables are age-standardized. Significance levels: *p<10%, **p<5%, ***p<1%.

We find no statistically significant differential impacts in child outcomes among children who were above or below the median of the same outcome measured at baseline (Table A27), though there is suggestive evidence that impacts in vocabulary are larger among children that exhibited lower language scores at baseline (measured with the PLS-IV scale), while impacts socioemotional outcomes are larger among those with higher socio-emotional outcomes at baseline.

Table A27: Heterogeneity of impact: baseline child outcomes

		NEP-B			NEP-I		
	Obs.	(1) Below Median	(2) Above Median	p-test (1)=(2)	(3) Below Median	(4) Above Median	p-test (3)=(4)
Vocabulary Index	2570	0.102	0.067	0.712	0.163**	0.065	0.313
		(0.068)	(0.067)		(0.069)	(0.067)	
Exec. Function Index	2890	-0.062	0.021	0.358	0.023	0.059	0.691
		(0.065)	(0.062)		(0.064)	(0.062)	
Socioemotional Index	2475	0.060	0.007	0.585	0.016	0.167**	0.120
		(0.069)	(0.067)		(0.069)	(0.068)	

Note: Each row reports estimates from a separate regression that includes interaction terms of intervention assignment with a dummy variable for whether the child was below or above the median outcome measured at baseline. Baseline vocabulary was measured with the PLS-IV test which measures receptive and expressive language for children from 6 to 72 months old. Baseline Executive Function was measured using the Dimensional Change Card Sort test for children older than 2 years old, and the A not B task for younger children. Socioemotional outcomes at baseline were measured with the Battelle Socio-Personal and the Child Behavior Checklist. Outcome variables are age-standardized. Significance levels: *p<10%, **p<5%, ***p<1%.

Heterogeneity by Other Maternal Outcomes

Finally, differential impacts in child outcomes by personality traits measured with the Big Five Personality test collected at follow-up are rarely significant and not following consistent patterns. For each personality trait, we compared mothers who were below or above the median standardized score (Tables A28-A32).

Table A28: Heterogeneity of impact: Caregiver's Conscientiousness (Big Five)

		NEP-B			NEP-I		
		(1)	(2)		(3)	(4)	
	Obs.	Below Median	Above Median	p-test (1)=(2)	Below Median	Above Median	p-test (3)=(4)
Vocabulary Index	2895	0.077	0.082	0.955	0.161**	0.062	0.286
		(0.070)	(0.059)		(0.069)	(0.060)	
Exec. Function Index	2895	-0.066	0.006	0.433	-0.004	0.064	0.457
		(0.070)	(0.058)		(0.069)	(0.060)	
Socioemotional Index	2492	-0.001	0.049	0.617	0.043	0.118*	0.444
		(0.075)	(0.063)		(0.073)	(0.064)	

Note: Each row reports estimates from a separate regression that includes interaction terms of intervention assignment with a dummy variable for whether the mother was below or above the median standardized Conscientiousness, measured with Big Five Personality test. Outcome variables are age-standardized. Significance levels: *p<10%, **p<5%, ***p<1%.

Table A29: Heterogeneity of impact: Caregiver's Openness (Big Five)

		NEP-B			NEP-I		
		(1)	(2)		(3)	(4)	
	Obs.	Below Median	Above Median	p-test (1)=(2)	Below Median	Above Median	p-test (3)=(4)
Vocabulary Index	2895	0.053	0.100	0.605	0.106	0.091	0.872
		(0.065)	(0.062)		(0.066)	(0.062)	
Exec. Function Index	2895	0.011	-0.054	0.467	0.126*	-0.036	0.077
		(0.064)	(0.062)		(0.066)	(0.062)	
Socioemotional Index	2492	0.126*	-0.059	0.054	0.163**	0.029	0.165
		(0.068)	(0.066)		(0.069)	(0.066)	

Note: Each row reports estimates from a separate regression that includes interaction terms of intervention assignment with a dummy variable for whether the mother was below or above the median standardized Openness, measured with Big Five Personality test. Outcome variables are age-standardized. Significance levels: *p<10%, **p<5%, ***p<1%.

Table A30: Heterogeneity of impact: Caregiver's Neuroticism (Big Five)

		NEP-B			NEP-I		
		(1)	(2)		(3)	(4)	
	Obs.	Below Median	Above Median	p-test (1)=(2)	Below Median	Above Median	p-test (3)=(4)
Vocabulary Index	2895	0.094	0.081	0.888	0.016	0.167***	0.109

		(0.071)	(0.059)		(0.070)	(0.060)	
Exec. Function Index	2895	-0.086	0.026	0.229	0.043	0.041	0.983
		(0.070)	(0.059)		(0.070)	(0.059)	
Socioemotional Index	2492	0.026	0.050	0.814	0.053	0.125*	0.473
		(0.076)	(0.063)		(0.074)	(0.064)	

Note: Each row reports estimates from a separate regression that includes interaction terms of intervention assignment with a dummy variable for whether the mother was below or above the median standardized Neuroticism, measured with Big Five Personality test. Outcome variables are age-standardized. Significance levels: *p<10%, **p<5%, ***p<1%.

Table A31: Heterogeneity of impact: Caregiver's Agreeableness (Big Five)

		NEP-B			NEP-I		
		(1)	(2)		(3)	(4)	
	Obs.	Below Median	Above Median	p-test (1)=(2)	Below Median	Above Median	p-test (3)=(4)
Vocabulary Index	2895	0.190***	-0.005	0.035	0.059	0.126**	0.471
		(0.070)	(0.059)		(0.070)	(0.060)	
Exec. Function Index	2895	-0.016	-0.027	0.902	0.008	0.058	0.585
		(0.069)	(0.059)		(0.070)	(0.059)	
Socioemotional Index	2492	0.056	0.010	0.640	0.042	0.121*	0.429
		(0.074)	(0.063)		(0.075)	(0.063)	

Note: Each row reports estimates from a separate regression that includes interaction terms of intervention assignment with a dummy variable for whether the mother was below or above the median standardized Agreeableness, measured with Big Five Personality test. Outcome variables are age-standardized. Significance levels: *p<10%, **p<5%, ***p<1%.

Table A32: Heterogeneity of impact: Caregiver's Extroversion (Big Five)

		NEP-B			NEP-I		
		(1)	(2)		(3)	(4)	
	Obs.	Below Median	Above Median	p-test (1)=(2)	Below Median	Above Median	p-test (3)=(4)
Vocabulary Index	2895	0.082	0.076	0.944	0.037	0.138**	0.288
		(0.075)	(0.056)		(0.074)	(0.057)	
Exec. Function Index	2895	-0.027	-0.021	0.946	0.147**	-0.026	0.065
		(0.074)	(0.056)		(0.073)	(0.057)	
Socioemotional Index	2492	0.165**	-0.047	0.036	0.181**	0.036	0.150
		(0.080)	(0.060)		(0.080)	(0.060)	

Note: Each row reports estimates from a separate regression that includes interaction terms of intervention assignment with a dummy variable for whether the mother was below or above the median standardized Extroversion, measured with Big Five Personality test. Outcome variables are age-standardized. Significance levels: *p<10%, **p<5%, ***p<1%

Table A33: Facilitator Background by Treatment Arm

TIPO	NEP-B	NEP-I	Total
1. Nurses	41.4%	27.1%	34.2%
2. Psychologists	19.6%	23.6%	21.6%
3. Educators	14.5%	39.0%	26.8%
4. Social Workers	24.5%	10.4%	17.4%
Health Workers (1+2)	61.0%	50.6%	55.8%

Table A34: Heterogeneity of impact: Facilitator Background

	NEP-B			NEP-I		
	(1) Health Worker	(2) Non health worker	p-test (1)=(2)	(3) Health Worker	(4) Non health worker	p-test (3)=(4)
Vocabulary Index	0.061	0.044	0.835	0.068	0.105*	0.632
	(0.056)	(0.066)		(0.060)	(0.060)	
Exec. Function Index	-0.007	-0.046	0.617	0.043	0.039	0.963
	(0.055)	(0.066)		(0.059)	(0.060)	
Socioemotional Index	-0.013	0.102	0.168	0.012	0.176***	0.044
	(0.059)	(0.071)		(0.063)	(0.064)	

Note: Each row reports estimates from a separate regression that includes interaction terms of intervention assignment with a dummy variable for whether the facilitator is a health worker or not. Outcome variables are age-standardized. Significance levels: *p<10%, **p<5%, ***p<1

Table A35: Heterogeneity of impact: Type of Health Center

	NEP-B			NEP-I		
	(1) General Health Center	(2) Family Health Center or Hospital	p-test (1)=(2)	(3) General Health Center	(4) Family Health Center or Hospital	p-test (3)=(4)
Vocabulary Index	0.065	0.070	0.963	0.038	0.115**	0.488
	(0.096)	(0.050)		(0.098)	(0.050)	
Exec. Function Index	-0.144	0.012	0.146	-0.080	0.074	0.158
	(0.096)	(0.050)		(0.097)	(0.050)	
Socioemotional Index	0.119	0.009	0.334	0.385***	0.014	0.001
	(0.102)	(0.053)		(0.102)	(0.053)	

Note: Each row reports estimates from a separate regression that includes interaction terms of intervention assignment with a dummy variable for whether the health center is a general type of center or a specialized type including family health center or small hospitals. Outcome variables are age-standardized. Significance levels: *p<10%, **p<5%, ***p<1

Appendix 6: Mediation Analysis with individual mediators

Table A36: Mediation analysis: Vocabulary Index

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
	Base	Home Index	PBC Nurturing Index	PBC Positive Discipline	Perceived Self-efficacy	PACOTIS	Perceived Friends Support	Significant Mediators
NEP_B	0.070	0.063	0.068	0.069	0.070	0.071	0.070	0.063
	(0.044)	(0.044)	(0.044)	(0.045)	(0.045)	(0.044)	(0.045)	(0.044)
NEP_I	0.099**	0.090**	0.096**	0.098**	0.096**	0.102**	0.099**	0.091**
	(0.045)	(0.045)	(0.045)	(0.045)	(0.045)	(0.045)	(0.045)	(0.045)
Home Index		0.093***						0.098***
		(0.020)						(0.021)
PBC Nurturing Index			0.032*					-0.013
			(0.020)					(0.022)
PBC Positive Discipline Index				0.006				
				(0.019)				
Perceived Self-efficacy					0.021			
					(0.019)			
PACOTIS						-0.037		
						(0.023)		
Perceived Friends Support							-0.005	
							(0.019)	
Observations	2895	2895	2895	2895	2893	2895	2895	2895
% Indirect Effect		**7.7%**	**2.5%**					**7.5%**
Confidence Intervals for the Joint Significance								
Lower Bound		0.01%	-0.01%					0.01%
Upper Bound		18.06%	8.00%					18.67%

Note: Each column reports estimates from a separate regression. Child outcome is internally age-standardization to the full sample (mean 0, SD=1). Estimates control for child sex and gender, and for health center's fixed effects. Column 1 presents the ITT outcomes as a benchmark (Tables 3 and 4). Columns 2-7 include one potential mediator at a time in the outcome equation. Column 8 include all significant mediators. The last row reports the total indirect effect in the child outcome that is attributable to intervention effects in mediators. Significance levels: *p<=10%, **p<=5%, ***p<=1%.

Table A37: Mediation analysis: Socioemotional Index

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
	Base	Home Index	PBC Nurturing Index	PBC Positive Discipline	Perceived Self-efficacy	PACOTIS	Perceived Friends Support	Significant Mediators
NEP_B	0.032	0.017	0.023	0.022	0.027	0.023	0.024	0.007
	(0.047)	(0.046)	(0.046)	(0.047)	(0.046)	(0.047)	(0.047)	(0.045)
NEP_I	0.094**	0.074	0.077*	0.077	0.070	0.080*	0.083*	0.048
	(0.047)	(0.046)	(0.046)	(0.047)	(0.046)	(0.047)	(0.047)	(0.045)
Home Index		0.245***						0.156***
		(0.021)						(0.022)
PBC Nurturing Index			0.223***					0.102***
			(0.021)					(0.023)
PBC Positive Discipline Index				0.160***				0.058***
				(0.021)				(0.021)
Perceived Self-efficacy					0.251***			0.158***
					(0.020)			(0.022)
PACOTIS						0.130***		0.035*
						(0.021)		(0.021)
Perceived Friends Support							0.108***	0.014
							(0.020)	(0.020)
Observations	2492	2492	2492	2492	2490	2492	2492	2492
% Indirect Effect		**22.0%**	**17.9%**	**17.1%**	**26.6%**	**14.2%**	**9.5%**	**50.4%**
Confidence Intervals for the Joint Significance								
Lower Bound		-1.54%	-3.45%	1.55%	1.85%	1.59%	-0.01%	24.18%
Upper Bound		46.19%	39.98%	34.40%	51.88%	28.34%	21.70%	78.36%

Note: Each column reports estimates from a separate regression. Child outcome is internally age-standardization to the full sample (mean 0, SD=1). Estimates control for child sex and gender, and for health center's fixed effects. Column 1 presents the ITT outcomes as a benchmark (Tables 3 and 4). Columns 2-7 include one potential mediator at a time in the outcome equation. Column 8 include all significant mediators. The last row reports the total indirect effect in the child outcome that is attributable to intervention effects in mediators. Significance levels: *p<=10%, **p<=5%, ***p<=1%.

Appendix 7: Cost-benefit Analysis

Rationale for Program Costs

Labor costs: Facilitators were paid an hourly rate of 6,180 CLP ($18.1 in 2022 US dollars)[19], regardless of the treatment arm, which includes preparing and delivering the session. For NEP-B, the budget considered each facilitator would deliver six two-hour sessions plus one hour of preparation across all sessions, for a total of 13 hours of work per facilitator. For NEP-I, the budget considered two extra sessions relative to NEP-B, for a total of 17 hours of work per facilitator. As a result, the total cost per facilitator (child) was $235.7 ($39.3) in NEP-B and $308.2 ($52.4) in NEP-I.

Materials: The annual budget for materials considered a unit cost of CLP 3,912 per child attended in the NEP-B program ($11.4 in 2022 US$ dollars). Since NEP-I facilitators were expected to conduct 8 sessions (2 additional sessions, 33% increase), the cost of materials per child in this arm was adjusted by the increase in sessions totalizing $14.7.

Training Costs:

1) ***NEP Intensivo (NEP-I):*** The total budget for training facilitators in this arm included the following items:
   - NEP Intensive Manual for facilitators (design, elaboration, video edition, and pilot of the sessions): $6,000 (US$2010 dollars)
   - One day of training of 162 facilitators in 8 groups (2 in North, 2 in South, and 4 in the Metropolitan area: $6,400. It includes two trainers per session.
   - One day of fieldwork follow-up: 162 NEP-I facilitators followed-up in 16 group sessions: 4 in North, 4 in South and 8 in Metropolitan area. It includes 1 trainer supervising each follow-up session.
   - On-line permanent follow-up: $1,600
   - Travel and subsistence Training Sessions: $1,600
   - Travel and Subsistence fieldwork follow-up: $2,000
   - Total training costs NEP-I in 2010 US$ dollars: $20,400
   - Total training costs NEP-I in 2022 US$ dollars: **$ 29,085** ($179.5 per facilitator; $6.0 per child assuming each facilitator delivers the program five times).

2) ***NEP Basico (NEP-B):*** Since we did not have access to the training costs for the standard program, we use the unit training costs for the NEP-I program to predict these costs under the following assumptions:
   - NEP-B manuals: The standard version of the program has five manuals, one for each of the session topics parents would choose. Since these manuals are much shorter than the NEP-I manual and do not include audiovisuals, we assume the cost of each NEP-B manual to be half the cost (US$3,000) of the NEP-I manual, for a total of $15,000.
   - Training sessions: NEP-B facilitators underwent a total of four 8-hour days of training. Assuming

[19] Exchange rate in 2010 1 US$= 486 CLP. Cost in 2022 obtained by adjusting to inflation at a 3% annual rate.

each day of NEP-B training had the same cost as the NEP-I training, then the 4-day training cost was $6,400x4=$25,600.

- Travel and Subsistence 4-day training sessions: $1,600x4= $6,400.
- Total training costs NEP-I in 2010 US$ dollars: $47000
- Total training costs NEP-I in 2022 US$ dollars: **$ 67,010** ($413.6 per facilitator; **$13.8** per child assuming each facilitator delivers the program five times).
- Total training cost per child NEP-B + NEP-I combined: **$19.8**

Rationale for Future Benefits and Costs

Table A38: Predicted Program Impacts on Adult Outcomes

	Cognitive (language)	Socioemotional	Total
NEP-I Impact (LATE)	0.425	0.381	
Assumptions on Returns (from NCDS data)			
College education	0.126	0.012	0.138
Earnings	0.180	0.029	0.209
Employment	0.029	0.018	0.047
Crime	-0.009	-0.018	-0.027
Depression	-0.025	-0.013	-0.038
Program returns			0.000
College education	0.053	0.004	0.058
Earnings	0.077	0.011	0.088
Employment	0.012	0.007	0.019
Crime	-0.004	-0.007	-0.011
Depression	-0.011	-0.005	-0.016

Note: Program return on a given adult outcome is the product between the IV impact on a given child outcome and the estimated association between the adult outcome and a 1 SD increase the child outcomes. Returns to cognitive and socioemotional skills in adult outcomes are obtained using data from wave 6 of the National Child Development Study (NCDS) of the United Kingdom, which follows a representative cohort of N=1,313 individuals. Returns are estimated using linear regressions of outcomes at age 42 on cognitive and socioemotional skills measured at age 7. Controls include years of mother and father education, number of siblings, white, female, region. There is one continuous outcome (log annual earnings), while college education, employed, crime (any dealings with the police), and depressed are binary outcomes.

Table A39: Summary of lifetime costs and benefits

Detailed inputs for calculation of lifetime costs and benefits	
Future schooling costs	
Predicted program impact on college attendance	0.058
Cost of college attendance per student in 2019 (age 18)	$9,817

PDV of the cost of college attendance per student	$5,766
Program impact on college costs per student	***$334***
Earnings	
Predicted program impact on wages	0.088
PDV of lifetime earnings per child (adjusted by LFP)	$95,287
PDV of gains in lifetime earnings per child	**$8,352**
Labor Force Participation	
Predicted program impacts on LFP	0.019
PDV of gains in lifetime earnings per child	**$2,565**
Crime Reduction	
Predicted program impact in crime reduction	-0.004
Cost per apprehended person in 2019	$33,490
PDV of gains in reduced crime per child over 10 years	**$1,236**
Mental Health	
Predicted Program impact on mental health	-0.016
Direct public mental health costs per person in 2021	$47
Direct Out-of-pocket mental health costs per person in 2021	$117
PDV of lifetime direct costs of mental health	$3,000
PDV of lifetime earnings losses due to mental health problems	$21,622
PDV of gains in mental health improvement	**$386**

Note: Program return on a given adult outcome is the product between the IV impact on a given child outcome and the estimated association between the adult outcome and a 1 SD increase the child outcomes. The total predicted return is the sum across the returns to cognitive and socioemotional outcomes. The estimated cost of college per student in Chile in 2019 obtained from public data reported by the Ministry of Education. Earnings and labor force participation rates profiles by age from a representative sample of individuals aged 25-64 from the 2019 *Encuesta Suplementaria de Ingresos (ESI).* Total costs of crime in Chile in 2014 reported by (Saens 2015). The number of apprehended individuals in 2019 is available at (CEAD 2019). Public mental health costs obtained from the 2021 national budget published by the Ministry of Health. Out-of-pocket mental health obtained from 2020 Encuesta de Proteccion Social 2020. Assumptions about labor costs of mental health obtained from (Ruiz-Tagle and Troncoso 2018).

Table A40: Cost-benefit Ratios with ITT Impacts

	Components of benefits			Sensitivity Analysis		
	(1) Earnings	(2) + LFP	(3) All costs and benefits	(4) 4 x Program costs	(5) + 50% of assumed returns	(6) + 5% discount rate
Costs						
Intervention costs	*$84*	*$84*	*$84*	*$337*	*$337*	*$337*
College attendance costs	$79	$79	*$79*	*$79*	*$40*	$28
Long-term societal costs	***$164***	***$164***	***$164***	***$416***	***$377***	***$365***
Benefits						
Gains in earnings	$1,994	$1,994	*$1,994*	*$1,994*	*$997*	$518
Gains due to increased LFP		$628	$628	*$628*	*$314*	$162

Gains in crime reduction			\$312	*\$312*	*\$156*	\$94
Gains in mental health			\$87	\$87	\$43	\$26
Long-term societal benefits	***\$1,994***	***\$2,622***	***\$3,021***	***\$3,021***	***\$1,510***	***\$800***
Cost-benefit Ratios	**12.2**	**16.0**	**18.5**	**7.3**	**4.0**	**2.2**

Note: The gain (or cost) in a specific adult outcome (e.g., earnings, labor force participation (LFP), crime reduction, mental health, college attendance) is the sum of the gains (or costs) induced by program ITT impacts in language and socioemotional development. The predicted gain (or cost) in a specific adult outcome induced by program improvements in a specific child outcome (e.g., language) is obtained as the product between the ITT intervention impact on the child outcome, the assumed correlation between the adult outcome and the child outcome, and the net benefit or cost per child, as described in the Appendix 7. All benefits and costs originally in Chilean pesos are converted to 2022 USD using the corresponding exchange rate to the year of the original data source and adjusted for inflation to January 2022. Present period total societal costs per child include direct program costs and the present value of long-term college attendance costs. Gains in adult outcomes are all in present values, where the discount rate is 3%.