AN EVALUATION OF TOSTAN’S REINFORCEMENT OF PARENTAL PRACTICES (RPP) PROGRAM

August, 2016

Dr. Ann Weber

Dr. Anne Fernald

Stanford University

The section on “Kaolack Caregiver’s Beliefs and Practices” was written by Drs. Ben Cislaghi and Diane Gillespie.

Acknowledgments

This report would not have been possible without the dedicated efforts of the Stanford-Tostan Evaluation Project (STEP) team members based in Dakar, Senegal, who were responsible for all of the data collection and coding: Serginho Donatien Badji, Abdoulaye Boussso, Mourtada Dem, Magor Dia, Yatma Diop, Sidy Giroux, Abdoulaye Kasse, Ibrahima Samba, Mamadou Thiam, Anna Vanderkooy, Ibrahima Ba, and Ndiaga Ndiaye. Jërejëf!

The contributions of several former and current staff and research assistants at the Stanford University Language Learning Lab were also absolutely integral to the success of this project, including support from Virginia Marchman, Melanie Ashland, Shamsi Soltani, Karl Pope, Kyle McDonald, Kat Adams, Ricardo Bion, and several of the Stanford students working in the lab.

We would especially like to thank our research collaborators, including Molly Melching and the Tostan staff, Ben Cislaghi currently at the London School of Hygiene and Tropical Medicine, Diane Gillespie at the University of Washington, and Ibrahima Giroux, former Tostan staff and part-time STEP team member.

We gratefully acknowledge the generous support of Madji Sock and her support staff at Dalberg Global Development Advisors: they took care of the STEP team in Dakar, managed the field work logistics, and offered solutions over every hurdle along the way. We could not have asked for a better partner.

Additionally, we are extremely grateful for the financial support (and patience!) provided by the William and Flora Hewlett Foundation for this important work.

Finally, we would like to thank the caregivers and children in Kaolack, Senegal, who participated in this study, and who welcomed the STEP team members into their communities with warmth, humor and generosity.
# Table of Contents

Executive Summary ......................................................................................................................... 3  
Introduction .................................................................................................................................. 4  
Section 1 – Tostan’s Program & the Social-Cultural Context .......................................................... 5  
  Tostan’s Reinforcement of Parenting Practices Program ................................................................. 5  
  Kaolack Caregiver’s Beliefs and Practices ....................................................................................... 6  
Section 2 – Impact Evaluation Design .............................................................................................. 9  
  Evaluation design & village selection .............................................................................................. 9  
  Staff recruitment & training .......................................................................................................... 9  
  Participant selection .................................................................................................................... 11  
  Data collection ........................................................................................................................... 12  
  Ethical Considerations ................................................................................................................ 12  
Section 3 – Measures ...................................................................................................................... 13  
  Measures of caregivers’ parenting skills ........................................................................................ 13  
  Measures of children’s language proficiency .............................................................................. 17  
  Other Measures ......................................................................................................................... 23  
Section 4 – Analyses ......................................................................................................................... 25  
  Raw Data Distributions .............................................................................................................. 25  
  Regression Estimates of Effect ..................................................................................................... 25  
Section 5 – Results .......................................................................................................................... 26  
  Baseline Comparison of RPP and Control Groups ..................................................................... 27  
  Changes in caregivers’ parenting skill ......................................................................................... 31  
  Changes in children’s language proficiency ............................................................................... 35  
Section 6 - Discussion .................................................................................................................... 39  
References ....................................................................................................................................... 41  

---

Stanford-Tostan Evaluation Project
Executive Summary

The Stanford-Tostan Evaluation Project (STEP) aimed to evaluate the efficacy of Tostan’s Reinforcement of Parental Practices (RPP) program in changing parenting skills and children’s language outcomes in the Wolof-speaking Kaolack region of Senegal. In this final project report, we describe the development, adaptation and testing of our measures, and present our results of the RPP impact after one and two years of program exposure. Specifically, we compare measures of caregiver behavior and child language in villages that received the RPP program with these same measures in villages that did not receive the RPP program, taking into account differences between the two groups that existed at baseline.

Overall, we found large and consistent changes in the caregivers living in the RPP villages after the first year of the program as compared to the comparison group at the post-program follow-up. Caregivers in the RPP villages showed greater improvements in their quantity and complexity of speech directed to children in one-on-one play sessions, in their knowledge of early child development, and in their attitude toward the use of physical punishment with children as compared to caregivers in the comparison villages. Most impressively, caregivers in the RPP villages nearly doubled their amount of child-directed speech in a structured play session as compared to caregivers from comparison villages who barely changed how much they talked with their child from one year to the next. Equally impressive was that children in the RPP group produced significantly more speech than did children in the comparison group during the play session, indicating a high degree of caregiver-child verbal engagement in the RPP group. In addition, we found small, but statistically significant, improvements in the caregiver-report measures of child language milestones achieved and expressive vocabulary.

Our findings from data collected two years after the start of the program showed that caregivers in the RPP villages continued to demonstrate a high-level of verbal engagement with their child when provided with one-on-one time to play. After the second program year, RPP caregivers had mostly retained their level of knowledge of child development and attitudes towards the use of corporal punishment with children. In addition, we found a long-term advantage in reported measures of child language abilities (i.e., language milestones and expressive vocabulary) for children whose caregivers had enrolled in the RPP program when the children were younger.

We were unable to detect an effect of the RPP program on children’s language processing efficiency as measured directly with the looking-while-listening (LWL) task in either follow-up period. In addition, all-day LENA™ audio recordings showed that children’s language experience during a typical day in program villages did not differ significantly from comparison villages after one or two years of the program.

Our study was subject to several limitations. While the Tostan RPP intervention was conducted in 200 villages across five regions of Senegal which differed in language and cultural traditions, our evaluation was limited to 24 Wolof villages in the Kaolack region. Thus we don’t know if our positive findings on the effect of the RPP program on amount of verbal engagement with young children can be generalized to the other language and cultural groups in which the program was conducted. In addition, Tostan had pre-selected villages to receive the RPP program, such that random assignment was not possible. But as the result of careful village sampling, key language measures at baseline across the two groups were...
found to be very comparable. An additional limitation was that although we relied on direct measures of both caregivers’ and children’s behaviors for the evaluation, we also used standard parent-report measures of children’s language skill that are inherently subject to over-reporting bias. However, the parent-report measures were positively correlated with child vocalizations obtained from all-day audio recordings, providing evidence of validity of the measures.

Overall, the results of the RPP evaluation represent an impressive achievement for Tostan. Senegalese mothers with no formal education, living in subsistence-level rural villages, were motivated to learn new ways of interacting with their young children, which in turn was associated with greater gains in their children’s language development. However, data captured in all-day LENA™ audio recordings suggested that children’s language experience during a typical day in the RPP villages did not differ from the comparison group. In fact, the amount of adult verbal engagement with children appeared to have decreased from one year to the next in both groups as the children grew older and more mobile. Therefore, a challenge remains for Tostan to identify how to increase the amount of rich adult verbal engagement that young children experience—by providing RPP training to all of the child’s caregivers—including older siblings and grandparents—and by improving the training to include new ways of incorporating child-directed talk while doing chores and activities throughout an ordinary day.

Introduction

In rural Senegal, very few children learn to read due to the poor quality of education and lack of family preparation for supporting children’s learning early in life. To address these barriers to children’s success and to the economic development of Senegal, Tostan developed a community-led approach to improving learner outcomes, the Reinforcement of Parental Practices (RPP) program. The initial goals of the RPP program were to enable caregivers to engage more effectively in verbal interactions with infants and young children, which in turn was expected to cause change in children’s early language skills, as a consequence of the richer cognitive stimulation they experienced. The longer-term goal of the RPP program is to show that these changes in caregiving practices and in children’s language skills in the first three years will have cascading consequences, leading to stronger cognitive skills and greater school success in later years.

The RPP program builds on scientific discoveries by developmental psychologists at Stanford and elsewhere on the crucial role of early language experience in cognitive development: Those children who hear more rich and varied language directed to them by caregivers develop vocabulary more rapidly (Rowe, 2012) and also become more efficient in real-time language processing (Weisleder & Fernald, 2013)—a critical skill that is predictive of later language development. The Stanford-Tostan Evaluation Project (STEP) provided rigorous evaluations of both caregivers’ and children’s behaviors and skills in the early years, based on the same fine-grained quantitative and qualitative measures used in laboratory research.

We assessed about 500 caregiver-child pairs in 24 Wolof-speaking communities in the Kaolack region, and compared longitudinal assessments of parenting practices and children’s language outcomes in a program group and a matched comparison group. Our goal in the STEP evaluation study was to
determine whether the RPP program improved parenting practices, and if so, whether more effective parent-child interaction in turn led to change in children’s early language development. In addition to adapting the standard questionnaire and interview measures commonly used in evaluation research, we included direct observations of caregiver-child interactions in a video-recorded play session, as well as all-day audio recordings of language heard in daily interactions, using the LENA™ speech recording and analysis technology (Ford, Baer, Xu, Yapan, & Gray, 2008). These measures generated extensive samples of caregivers’ and children’s speech in Wolof.

With these rich data, the STEP study had four overarching goals: (1) to evaluate the efficacy of the RPP program in changing parenting skills and children’s early language outcomes; (2) to provide the foundation for a longitudinal follow-up of these children at school age; (3) to develop robust testing procedures that yield sensitive continuous measures of early language development suitable for use in assessing effects of parenting programs in field conditions; and (4) to make a unique and valuable contribution to basic research and theory on early language growth, by extending new findings on the critical developmental role of early language experience to parents and children in an African culture.

The following report summarizes the STEP activities and results from late 2012 through mid-2016. In the first section, we describe the Tostan parenting program and findings from qualitative research on the social-cultural context of the communities where the program took place. In the second section, we present the details of the evaluation study design, the village and participant sampling procedures, and the demographics of our study sample. In the third section, we describe our measurement instruments and the process we used to adapt the instruments for the local Wolof context. In addition, we present pilot results from the assessments to demonstrate that we used reliable and culturally relevant measures of important aspects of children’s language development for Kaolack. In the fourth section, we describe the analytical methods. In the fifth section, we present our main findings, including the estimated effects of the program on both caregivers and children after one and two years of the program. We also demonstrate in this section that the sample characteristics are well-balanced between the RPP and comparison groups. In the final section, we discuss the results and the implications for the future success of Tostan’s parenting education program.

Section 1 – Tostan’s Program & the Social-Cultural Context

Tostan’s Reinforcement of Parenting Practices Program

Tostan is a respected NGO founded and based in Senegal since 1991 and now operating in six West African countries. Over 2000 rural communities have participated in Tostan’s Community Empowerment Program (CEP), a three-year human rights-based education program (Cislaghi, Gillespie & Mackie, 2016; Gillespie & Melching, 2010). The CEP engages communities to set a vision for their future while learning about democracy, human rights, hygiene, health, literacy and project management, and to decide which traditional practices help or hinder their development and well-being. In 2011, Tostan learned that caregivers living in many Senegalese villages reported that they avoided making eye contact or talking with their infants for fear that the baby would be possessed by evil spirits, a dire outcome...
thought to have fatal consequences (Zeitlin, 2011). Parents also reported that people are called “crazy” if they talk to a baby, because “nobody is there”. Concerned that low reading rates in the formal school in Senegal might in fact be related to such beliefs which hinder parent-child interaction early in life, Tostan developed the Reinforcement of Parental Practices (RPP) program, an early childhood development curriculum designed to follow and reinforce the three-year CEP. Drawing on scientific discoveries about the crucial role of early language experience in cognitive development and later school achievement, the initial goals of the RPP program were to enable caregivers to engage more effectively in verbal interactions with their infants, providing richer cognitive stimulation to build a strong foundation for their children’s learning skills.

With a focus on caregiving for children from 0-6 years, the RPP program in 2013 was comprised of 43 group sessions and bi-monthly home-visits over a 9-10 month period, conducted by a Tostan facilitator who lived in the community. Topics in the early sessions included the human rights of the child, the importance of the human brain, and scientific evidence on the influence of parenting practices on children’s language and cognitive growth. Since most participants had no experience reading to children, Tostan developed a set of 15 colorful children’s books in three national languages, and showed caregivers how to share them with children. Later sessions focused on how parents could help their children succeed in school – for example, by talking frequently to them, telling stories, and teaching them numbers, as well as engaging with teachers and monitoring children’s school progress. During the sessions, participants joined in games, role play, and other group activities. They also reflected on both the beneficial and potentially harmful aspects of traditional child care practices. During home visits, the Tostan facilitator reinforced what was taught in class, coaching caregivers as they practiced new activities with their children. Participants were encouraged to share their new knowledge with siblings, other relatives, neighbors and surrounding villages through Tostan’s model of “organized diffusion”, with the goal of increasing the impact of the RPP program and reducing the potential for sanctions by others.

The RPP program was continued for an additional year, throughout 2014. In the second year, Tostan focused on two major components: a social mobilization strategy and a community ownership and sustainability strategy. During this second phase of the program, they worked closely with religious leaders and engaged fathers in order to raise awareness and support for the important work that women are performing in raising their children. For the sustainability strategy, Tostan trained the most engaged caregivers in their program as “expert caregivers” and both the Tostan facilitator and the new “experts” continued to perform home visits during 2014.

**Kaolack Caregiver’s Beliefs and Practices**

Prior to the evaluation of Tostan’s parenting program, we contracted with Dr. Benjamino Cislaghi (then Tostan’s Director of Monitoring and Evaluation) of the London School of Tropical Medicine and Hygiene and Dr. Diane Gillespie (a member of Tostan’s board of directors) of the University of Washington to perform an assessment of the traditional belief systems shared by caregivers in the Kaolack region of rural Senegal, where the evaluation was to be performed. We were interested in beliefs and practices that might serve to encourage or discourage particular caregiving behaviors relevant to the RPP
program. In this study, Drs. Cislaghi and Gillespie were following up on informal reports about the power of such beliefs on parenting practices shared by Tostan researcher Dr. Marian Zeitlin (Zeitlin, 2011). Zeitlin reported wide-spread beliefs in a spirit world, in the existence of evil spirits, and in the need to protect one’s self from these spirits. Through semi-structured interviews and focus group discussions, they explored common parenting practices and the meanings that parents gave to them. Parents were asked specifically about the spirit world and the connection of that world to parental interactions with babies. For example, they asked: How widespread was knowledge about the spirit world and how did our research participants describe the effects or potential effects of these spirits on babies? How did they talk to their infants and how would significant others in their community judge different kinds of talking to babies?

Sixty-seven adult caregivers from three different Wolof-speaking communities in the Kaolack region participated in the qualitative research. All three communities formerly participated in Tostan’s Community Empowerment Program and were eligible to participate in the RPP program. Two Senegalese Wolof-speaking interviewers, one man and one woman, conducted the interviews and focus groups. They interviewed a total of 30 participants (with the male interviewer interviewing fathers and the female interviewer mothers). All interviewees had at least one child under the age of 3 years who they cared for daily, and most were caring for two or more young children. Eighteen had lost babies or infants. The remaining 37 participants took part in 6 gender-segregated focus groups of 5-7 participants each, facilitated by an interviewer of the same gender. Data were transcribed and translated by a local Wolof-speaking interpreter, who, along with the interviewers, helped with the cultural interpretation of participants’ responses. The data were first coded and then gathered into themes. Their results, reported below, include four main themes that emerged from the analysis.

Beliefs in the existence of a spirit world. The first theme that emerged was the widespread belief in the existence of a spirit world. Spirits, especially bad spirits, entered the real or physical world and caused harm, both to parents and their babies. For example, parents gave numerous examples of sickness and death in infants associated with spirits. A few parents expressed skepticism about the spirits, explaining that such a worldview was traditional and was inconsistent both with Islam and with modern medical practices.

Protection of children. Given this belief in spirits, the second theme that emerged was the need for child protection. Good parents took actions to guard against evil spirits during pregnancy and after birth. These practices were understood as beneficial. During and after pregnancy, for example, evil spirits looked for opportunities or “openings” to enter the pregnant woman or the newborn infant. Without such protection, the spirits would replace the real baby with a spirit baby who would die shortly after birth or fail to speak or walk. As a result, mothers took a number of precautions. For example, when pregnant, mothers remained in the house at certain times of the day and urinated outside in special places while carefully covering themselves with cloth. After birth and during the baby’s first week of life, the baby and the mother are restricted to a room in the house. To keep spirits away, several described placing a knife under the baby’s bed and putting an amulet on the baby’s wrist. Parents described remaining on guard against spirits during the first two years of a baby’s life.
Resulting problems from evil spirits were considered to be the parent’s fault, typically the mother’s. All participants knew about, and most had gone to, marabouts, Islamic religious leaders and teachers, many of whom practice animist rituals and ceremonies, such as making amulets. Good parents took their babies to marabouts, especially if they thought that they had been harmed by evil spirits. Taken as a whole, the protective practices were visible to others and significantly restricted the mothers’ and babies’ engagement in and with the world.

The interviewers asked participants specifically if talking to babies and looking at babies in the eyes would be dangerous and attract evil spirits. A few said that their grandparents believed that one should not look the baby in the eye and a few said that the gaze needed to be broken and not sustained, but the overwhelming majority of our participants did not say it was dangerous to look the baby in the eye, and one said that it would develop the baby’s intelligence. In the participants’ descriptions of their verbal interaction with their babies, they did not say that speaking to the baby could be dangerous in terms of inviting evil spirits. Rather their quite limited verbal engagement resulted from their beliefs about intellectual development, namely that babies were not able to understand language or what they were seeing in the world and so talking to them was meaningless, some said “crazy.”

**Beliefs about child development.** The third theme was the limited capacity of the infant to learn. All participating parents said that they had not talked to their babies in utero, with some expressing surprise that some parents would do that. After birth, the typical verbal interaction included single word or short imperatives, such as “suck” or “don’t cry,” or making nonverbal sounds, such as “pssst” or tongue clicking. Many parents said that babies did not start learning until much later, with ages of onset ranging from 3-7 years.

**Engagement of fathers.** The fourth theme that emerged was the active engagement of the fathers in the life of their babies. Several described their fear of talking and playing with their babies too much as they worried that they would be overly attached to them and their babies would not be able to withstand the separation that would come from the father’s absence or his possible death. Fathers, more than mothers, advocated for better infrastructure for the education of their children.

These four themes have important implications for the implementation of Tostan’s Reinforcement of Parental Practices. First, the constrictive practices related to pregnancy and parenting of babies are governed by beliefs about the harm that could come to their infants from the supernatural world if those practices are not followed – a complex but rational framework for infant care that needs to be taken into account. And second, the limited verbal interactions with infants did not appear to be the result of supernatural beliefs but rather a lack of scientific information about the timing of brain development and children’s capacity to learn.
Section 2 – Impact Evaluation Design

Evaluation design & village selection

In order to evaluate the RPP program, we designed a rigorous ‘proof-of-concept’ study with the central goal of testing whether caregivers’ behavior – in particular their tendency to engage in verbal interactions with their young children – actually changed as a result of participation in the program, and if so, whether these changes influenced children’s language learning. Although Tostan offered the RPP program in 3 of the 11 national languages spoken in different regions of Senegal, we focused on just one - Wolof, the most widely spoken language. Given the substantial demands of translating and validating measures specifically designed to assess parents’ language production and children’s language learning in Wolof, it was not feasible to develop comparable materials in other languages at the same time.

Tostan’s preparations for launching the RPP program in 200 Wolof-, Pulaar-, and Mandinka-speaking villages in five regions of Senegal were well underway before the evaluation study was designed. Therefore, random assignment of the program to villages was not possible. Faced with this constraint, we designed a quasi-experimental study with assessments pre- and post- the program in both RPP and comparison (or control) communities, and used a sampling strategy to minimize initial differences between the two groups. For better comparability between groups, we restricted the village sampling to a single region of Senegal, Kaolack, in which Wolof is the primary language spoken.

In addition, only villages that had formerly participated in Tostan’s Community Empowerment Program (CEP) were eligible to host the RPP program. Therefore, we further restricted the sampling frame of villages to 90 communities that had received the CEP. Forty of these communities had already been selected by Tostan to receive the RPP, and the remaining 50 communities were eligible to receive the RPP at a later date. Using data from Tostan in combination with GPS data on the size and location of these villages, we restricted the sampling frame to 20 of the 40 program sites and 17 of the 50 delayed-program sites. All of these villages were similarly isolated, large enough to meet our within-village sample size requirements, and estimated to have more than 90% Wolof-speaking households. We then performed systematic random sampling ordered by village size to select 12 program and 12 comparison sites, for a total of 24 villages included in the evaluation study. Two program sites were replaced during fieldwork with the next largest village in the list - one because the predominant language spoken was Bambara, and the other because the field team was unable to find an adequate testing space. Two comparison sites were also replaced - one because the village leaders refused to participate, and the other because the village had extremely unreliable age information for the children (e.g., date of birth was missing on most health cards). Nine of the 12 program villages had participated in the CEP as recently as 2011, whereas all but one of the comparison villages had completed the CEP prior to 2011.

Staff recruitment & training

Stanford contracted with Dalberg Global Development Advisors, a consulting firm based in Dakar, to recruit the field team members and to manage the fieldwork logistics. Two teams of four people each were hired and each team included a team leader/chef d’équipe, an interviewer, a child examiner, and a technical assistant, with the following responsibilities:
• The team leader was responsible for team management and supervision, logistics, data verification, communications with project staff, and the security of project personnel and equipment in the field. As such, in each village, the team leader introduced his members to the local authorities, including the village chief, and explained the project to them. He arranged room and board in each village, controlled the quality of the data by verifying all questionnaires at the end of the day, and ensured that all collected data were protected and transferred according to instructions. In addition to making sure that each member was doing his job well, the team leader acted as a daily point of contact with project staff at Dalberg and Stanford. This involved writing a daily report concerning the day’s activities, in which the team leader reported the number of questionnaires and tests completed, challenges faced by the team, issues to be resolved, and suggested solutions.

• The Interviewer administered the survey questionnaires, communicated with interview subjects and the team leader to coordinate testing sessions, and assisted during anthropometric measurements.

• The Child Examiner administered the primary caregiver questionnaire, distributed the LENA™ devices to target children, and assisted the technical assistant in the testing set-up.

• The Technical Assistant was responsible for the preparation of the test site, the use and maintenance of all equipment, and the recording, saving, and organization of all collected data.

STEP representatives at Dalberg, Tostan, and Stanford collaboratively trained the field teams from December 2012 to mid-January 2013 prior to the baseline field work. In-office training involved knowledge-building; preparation for the use of robust research methods and tools and anthropometric measures; and extensive discussion to improve and adapt chosen research tools to the cultural, linguistic, and geographical context of rural Senegal. Because Wolof is not taught as a written language in Senegalese schools, a significant challenge to the training component of the project involved teaching the field team members to read and write fluently in Wolof, a necessary skill for the successful administration of questionnaires and tests in Wolof. The field team was trained for one week about the rules of written Wolof and reviewed each questionnaire in detail. A few field team members had had prior experience reading and writing in Wolof and were thus able to share expertise with their colleagues during small-group questionnaire practice sessions.

Full-scale pre-testing in three non-study, Wolof-speaking villages in the region of Thies followed in-office training, allowing the teams to practice, adapt, and verify field research procedures. Stanford researchers were available for several weeks of the training period and pilot testing to provide instruction and direct supervision in the use of testing methods unfamiliar to the field teams. Both field teams, along with several STEP staff members from Dalberg, Tostan, and Stanford, went together to the first village, where field teams practiced administering questionnaires and tests. Working closely together for a week in the same village facilitated knowledge-exchange between the two field teams, who further harmonized testing procedures and signaled incoherencies in questionnaires. After several days of meetings in Dakar to update questionnaires and clarify testing procedures, the two field teams conducted an additional week of pre-testing in two separate non-study, Wolof-speaking villages in the region of Thies. STEP supervisory staff members and Stanford mentors were present in both villages,
confirming that field teams were consistent in their procedures throughout questionnaire administration and child testing. Pre-testing in Thies was repeated each year prior to the field work in Kaolack.

**Participant selection**

In each village, the team enrolled a median of 21 caregiver-child pairs (range: 12-30 pairs) from a list provided by village leaders of age-eligible children in two cohorts: children 4-19 months who were mostly preverbal, and children 20-31 months who were expected to be talking. The process of recruiting the child-caregiver pairs into our study was a multi-step process. Prior to the team’s arrival in a study village, a Tostan staff member visited each community to present the STEP study details to the village chief and community program coordinator. The Tostan employee started with a preliminary list of possible candidate participants obtained from a survey performed by Tostan as part of their village recruitment process into the RPP program. Upon arrival, the STEP team sought the village leaders’ cooperation in identifying any additional age-eligible children in the village.

All children in the older cohort were selected for the study. Since the list of younger children was longer, systematic random sampling was performed to meet our target of 10 children per cohort per village. Seventy of the caregivers recruited across the 24 villages were unable to enroll in the evaluation: 41 were travelling; 9 declined; and 23 could not attend the testing sessions. To meet our target sample size in each village, caregiver-child pairs living in the same polygynous household were recruited to participate, although no caregiver was represented twice. Children were excluded if they were reported by the caregiver to be bilingual (i.e., were spoken to more than 10% of

*Figure 1* illustrates participation by program assignment, age cohort and survey phase. Children in the control group are shown in orange, and those in the RPP group in blue. The darker shades represent the younger cohort, which was 4-19 m of age at baseline in 2013, 16-31 m at phase 2 in 2014 and 28-43 m at phase 3 in 2015. The lighter shades represent the older cohort, which was 20-31 m at baseline and 32-43 m at phase 2. Only the younger cohort from the RPP villages (darker blue) were included in phase 3.
the time in a language other than Wolof), or to have a serious developmental delay or a hearing, speech or vision impairment. If a mother had twins, only one was included. We chose as primary caregiver the individual who spent the most daylight hours responsible for the child and thus had the greatest opportunity to influence the child’s language development. Almost all primary caregivers (93.9%) were the target child’s mother.

At the phase 2 follow-up (early 2014), we found that not all of our RPP caregivers had both taken classes and received home visits in 2013. To address the concern that low participation would compromise our effect estimates, we augmented our analytic sample by returning to the RPP villages in early 2015 for a third phase of data collection (2 years after the start of the program). Using phase 2 data, we targeted 94 children who had been less than 20 months at baseline (young cohort), and whose caregivers had not changed since baseline and had participated in both the RPP classes and home visits in 2013, during the first year of the program.

Data collection

Extensive survey, video and audio data were collected for over 500 participants by two field teams at baseline in 2013 (phase 1) and one year later in 2014 (phase 2) in both RPP and comparison communities. A third data collection round (phase 3) was performed in 2015, two years after the baseline, but only in the 12 RPP villages, and only the younger cohort of children were assessed.

At phases 1 and 2, the two teams kept to a rigorous 10-week schedule of work and travel. A single team spent 5 weeks in the field for phase 3. Although the teams were given a warm welcome and received generous support in the villages, they faced strenuous daily challenges, including absence of electricity and clean water, poor sanitation, exposure to extreme heat, and the need to re-schedule data collection around unpredictable events. In each village, the teams set up a common testing space with adequate natural light for video recordings, but away from excessive noise and curious onlookers. The electronic equipment selected for these conditions operated on batteries which were recharged with a gas-powered generator every night. The teams remained in each village for 4 to 6 days, sleeping in tents with mosquito netting on the floor of a space provided by village leaders, before traveling on to the next village. Since internet with large bandwidth was not readily available, audio and video data were transferred on USB keys to a Dalberg liaison on travel days, to avoid loss from theft or other misfortune.

We used paper questionnaires at baseline for survey data collection with double data entry in Dakar. A major improvement implemented in the phase 2 follow-up was the replacement of all paper questionnaires with electronic versions. The survey software, Suverybe, was used for the design of the Computer Assisted Personal Interviewing (CAPI) questionnaires, data collection in the field, and export of the analysis-ready data. Survey data was transferred routinely by the team in Kaolack via email directly to the Stanford lab, obviating the need to perform multiple manual data entries at a later date.

Ethical Considerations

Approval for this research was obtained from Stanford University’s Institutional Review Board. Study protocols were also presented to the National Director of the National Agency for Applied Scientific
Sections 3 – Measures

Data for three broad categories of measures were collected to evaluate the effectiveness of the Tostan RPP program: (1) Measures of caregivers’ parenting skills, including direct observations of caregiver-child interaction in video-recorded play sessions and all-day audio recordings of children’s language environment using the LENA™ speech and analysis technology, as well as questionnaire data on caregivers knowledge of child development; (2) Measures of children’s language proficiency, including direct measures of children’s speech production from the video and audio recordings, direct assessment of language processing efficiency using a looking-while-listening task, as well as caregiver-report questionnaires on milestones in language growth and expressive vocabulary size; (3) Extensive survey data on potential confounding variables related to child, caregiver, household, and village characteristics.

Measures of caregivers’ parenting skills

Caregiver’s speech to the child in a structured play session. Hundreds of studies on early child development demonstrate the importance of rich non-verbal communication between parents and infants in the early months of life to establish a strong foundation for language as well as for social attachment. The value of parenting behaviors such as making eye contact with the baby and responding contingently to the infant’s expressions and vocalizations have been well documented in research on parents and infants in middle-class populations in the U.S., Europe, and Asia. And parents’ skill in establishing communicative routines with their infant has been linked to later success in language learning. However, in other populations, including agrarian societies in Central America and Africa, as well as in some minority groups in the U.S. and Europe, these strategies for early communication with infants are much less commonly used (LeVine et al., 1996; Shneidman & Goldin-Meadow, 2012). Such differences are often simply dismissed as reflecting “cultural preferences” for different parenting styles. However, increasingly there is concern that such reduced levels of parent-infant engagement are costly to the developing child, as more and more evidence becomes available that age-appropriate social and cognitive stimulation starting at birth are critical for optimal brain development.

One goal in our study was to provide data based on direct observation of caregiving skills by parents and other primary caregivers of infants and young children, rather than relying on the less sensitive self-report measures that are commonly used in such research.

Observation (OBS) Procedure: All children enrolled in our study were video-recorded in a play session with their primary caregiver. Video recording took place in the same testing locale identified in each village that was as private, quiet, and well-lit as possible. The caregiver and child pair sat on a floor-mat with familiar toys and household objects provided for the child to play with. A microphone was
attached to the caregiver’s clothing and a video camera set up on a tripod about 5 to 6 feet from the 
pair. The caregiver was asked to behave with her child just as she would at home, and was then left 
undisturbed in the room for 15 min in the pilot phase and at phase 1, and for 8 min at phases 2 and 3.

OBS Piloting: The OBS procedure used for naturalistic observation was piloted in five Senegalese villages 
during the summer of 2012, where more than 30 mothers and toddlers were recorded. Mothers and 
children appeared to be quite comfortable in the situation. The amount of actual interaction occurring 
over 15 min varied significantly among mother-child dyads in this small sample. For example, some 
mothers interacted enthusiastically with their children using varied and continuous speech, engaged in 
turn-taking, and often shared eye 
contact with the child. At the 
other extreme, some mothers 
vocalized only a few times in 15 
min using only one-word 
utterances. Our experience pilot-
testing the video sessions made 
us confident that the ‘demand 
characteristics’ of this situation 
(i.e., making the mother feel she 
should ‘perform’ for the camera) 
would not be a problem in the 
baseline study.

OBS Transcriptions & Word 
Counts: The middle five minutes 
of each recording was transcribed 
word-for-word in Wolof by one of 
the STEP field team members. 
We ignored the first 5 min of the 
video in order to give the 
caregiver a chance to settle into 
the situation. Caregiver and child 
speech was transcribed and then 
processed using the software 
package CLAN (Brian, 2000) to 
 obtain two measures of caregiver speech directed to the child in 5 min: total number of caregiver words 
and the mean length of caregiver utterances (MLU), calculated as the total number of words divided by 
the number of utterances. MLU is used here as a proxy for the complexity of caregiver speech to the 
child.

For the transcriptions, the team members were trained in Wolof orthography, as well as transcription 
protocols by a trained Wolof linguist. At baseline, every completed transcription was reviewed and 
corrected by one of our gold standard transcribers for consistency of spelling and parsing of utterances.
At both follow-ups, the transcription training was repeated. Twenty percent of the follow-up videos transcribed (94/469) were double-transcribed by a second person, and the word and utterance counts for both caregiver and child compared. We considered that the transcriptions agreed if the difference between the two transcribers’ counts was within 10% of the average of their two counts. If the average of the counts was 0-2 words or utterances, we categorized the transcriptions as in agreement. Inter-transcriber comparisons that differed by more than 10% were reviewed by the transcribers together to generate a final corrected transcription. Agreement for two transcribers was 98% of caregiver word counts, 84% of caregiver utterance counts, 53% of child word counts, and 73% of child utterance counts. Disagreement was most likely to occur for child word counts because young children’s utterances were frequently unintelligible and were sometimes difficult to hear and parse clearly. The number of caregiver utterances was not as reliable a measure as the number of caregiver words because of the ambiguity inherent in identifying utterance boundaries. Based on these reliability results, we used child utterances and caregiver words in our main analyses. We did not find a strong pattern of disagreement in counts as a function of the quantity of speech.

LENA™ audio-recordings of children’s language environment: While the 5 min naturalistic observation of caregivers and children captured a short sample of speech to the child when they were alone and without distractions, this may be a very atypical situation for children in Kaolack, who may spend most of their time interacting with other children rather than with adults. For this reason, we also used a technology called LENA™ (Language ENvironment Analysis) (http://www.lenafoundation.org), which enables unobtrusive all-day recordings of a child’s language environment without an observer present. At all phases, we collected rich language data on a typical “day-in-the-life” of Senegalese children as they interacted with many different people in the course of daily activities in the village. Although mother-child language interactions may be minimal in some cases, there are many other potential sources of language available to children, which LENA™ enabled us to sample.

LENA Procedure: All-day recordings were made using the LENA™ digital recorder, which was placed in the chest pocket of specially designed clothing worn by the child. The recorder has a single microphone located 7-10 cm from the child’s mouth, designed to record the audio environment within a 1.2-1.8 m radius - from the perspective of the child - for up to 16 hrs (16 kHz sampling rate). The audio data were transferred at the end of the day to a computer and then a USB key that was transferred to Stanford for later analysis.

LENA Piloting: We piloted the LENA™ technology in Senegal in 2012, and successfully obtained 8-hour recordings from 10 toddlers that convinced us that using LENA™ for the evaluation was feasible.

LENA Coding: The audio data were processed with LENA™ analysis software, which incorporates speech recognition algorithms to differentiate speech-related sounds from environmental background noise (Ford et al., 2008). This software yields automated measures of adult word counts (AWC), which is the number of words a child hears from adult males or females during the recording. Although the AWC does not distinguish child-directed from adult-directed speech, the software provides a breakdown of speech by “vocalization activity blocks.” Each block represents periods of speech separated by silence, categorized by who initiated the speech (e.g. adult female, adult male, target child, or other child). To
assess possible caregiver speech to the child, we used female AWC from blocks categorized as “adult female-initiated speech to target child”, since all primary caregivers were female. Because length of recordings varied, total counts were converted to counts per hour. (See the section on measures of children’s language proficiency for child-based measures obtained from the LENA™.)

**Caregiver Knowledge**: Basic caregiver knowledge about child development was captured with close-ended survey questions. The questions were developed in collaboration with Tostan to closely reflect the key concepts taught as part of the RPP module and consisted of a set of 15 statements read aloud to caregivers, such as “A baby’s brain starts to develop during pregnancy,” to which caregivers reported if they strongly agreed, agreed, disagreed or strongly disagreed. Responses to each item were scored on a 0- to 3-point scale, from low to high knowledge, and summed to create an index score of knowledge.

In pilot tests with 25 women in 3 villages in the region of Thies, we found that mothers were fairly well informed about child development. Overall, the 15 items were reliable in the pilot study with high coefficients for internal consistency (Cronbach’s alpha = 0.79). Although the coefficient for internal consistency was low at baseline (Cronbach’s alpha = .34), the knowledge indicator at baseline is used solely as a control variable in the regressions and did not differ statistically between the program and comparison groups. The interviewers reported that caregivers had difficulty with the Likert scale and recommended dividing the question into two parts: first asking if they agreed or disagreed, then if they agreed strongly or not (or disagreed strongly or not). We adopted this procedure at follow-up, and the coefficient for internal consistency of the 15 items was much improved at .76.

**Attitude toward discipline**: An attitude towards discipline scale from a UNICEF survey was adapted and translated for use in the STEP study. The scale consisted of an introductory statement that adults use different forms of instructing their child on good ways to behave or to deal with problem behaviors. The primary caregiver was then asked whether a parent had the right to hit or beat the child in each of 8 different situations (e.g., if the child is in danger of hurting herself). Responses to each item were scored on a 0- to 3-point scale, from strongly agree to strongly disagree, and summed to create an index score of attitude toward discipline (Cronbach's alpha = 0.78 at baseline and 0.89 at phase 2).

**Social Norms**: In addition to the qualitative research discussed in Section 1, we developed a quantitative module to assess traditional caregiver beliefs and social norms around infant care in rural Senegal. We started by identifying two main caregiver behaviors of interest based on the RPP program content and the work of Dr. Marian Zeitlin: talk to baby at an age-appropriate level (e.g., being responsive to child’s interest and responses, naming and elaborating on objects that child is interested in/looking at); and direct eye contact or mutual gaze between caregiver and baby. Our questions were aimed to assess social norms, or expectations that the community holds about the behavior of its members and are held in place by sanctions or rewards if the behavior is followed.

We initially planned to use videos shown on tablets to obtain the caregiver’s opinions on what she saw, without actually naming the behavior we were interested in (to avoid influencing her response). Four scenarios were recorded and tested in a pilot study: (1) caregiver and baby with eye contact and a caregiver “teaching” style of speech; (2) no eye contact and a “teaching” style of speech; (3) with eye
contact and caregiver giving command directives only (e.g., “stop that, don’t touch that”); and (4) no eye contact and command directives only. All scenes depicted the caregiver sitting with a baby on a mat where they could interact comfortably. The caregiver was holding or touching the baby in a natural way and was kept as constant as possible across the scenarios.

In the pilot study, five mothers were shown the two videos with eye contact and eight mothers were shown the two videos without eye contact. After viewing a video, each participant was asked if she behaves in the same way with her baby, if she thinks she should behave that way, whether other mothers would say she should behave that way, whether other moms behave that way, how the people she respects would say she should behave, and whether she approves or disapproves of the mother’s behavior in the video. In most cases, the pilot participants said they did what was shown in the videos multiple times a day, regardless of the video shown.

Respondents did not judge the eye-contact videos as different from the no-eye-contact videos. The respondents’ attention was on the type of talk - which they perceived as different, although they thought it important to do both. Most of the mothers we interviewed said it was important to talk to babies in a teaching way because it helps the babies to develop.

Based on the above results, we dropped the video approach for the social norms module. However, during the pilot study, we asked open-ended questions about beliefs and fears of looking a baby in the eyes or talking to babies. We used these responses to re-work the survey module (without videos) into two sets of similar questions around talk to babies and looking babies in the eyes, but where we named the behaviors of interest. This module was administered at baseline, but subsequently dropped due to lack of variability in the responses (data not shown).

Measures of children’s language proficiency

**Observed measures of children’s vocalizations.** To compare the amount of child vocalization between the RPP and comparison groups, and over time, we used the same play session transcripts and all-day LENA™ audio recordings described for caregivers. Since accurate assessment of number of words was problematic due to difficulty of transcribing specific words for the younger children, the measure of child language production in the play session was the total number of utterances produced in 5 min. From the LENA™ recordings, we obtained automated estimates of child vocalizations and conversational turns. A child vocalization was counted when child speech of any length was surrounded by > 300 ms of silence or other sound that was not child speech. A conversational turn was counted when a child vocalized and an adult responded within 5 s – or an adult spoke and a child responded. As with adult word counts, total child vocalization and turn counts were converted to counts per hour.

In order to test the reliability of the automated LENA™ all-day measures for Wolof-speaking children, we listened to and transcribed 1-hour samples from 41 of the children’s recordings at baseline. The correlation between the number of child utterances transcribed in the play session and the two automated all-day measures were good at .64 for child vocalizations and .65 for conversational turns (p-value < .001 for both).
Looking-While-Listening (LWL) procedure: In our research on early language learning over the past 15 years, we have developed powerful experimental methods for studying children's efficiency in interpreting language in real time. The “looking-while-listening” (LWL) procedure tracks infants' eye movements as they look at a pair of pictures (one target and one distractor image) while listening to speech, yielding precise measures of processing skill at millisecond-level resolution. Our first discovery in the U.S. was that processing efficiency at 24 months is highly correlated with early vocabulary development. Infants who are faster to recognize familiar words also learn new words more easily, resulting in faster vocabulary growth. After following a group of English-learning infants from 15 to 24 months of age, we observed the same children six years later as they entered 3rd grade. The LWL measures of real-time processing efficiency at 24 months were strongly predictive of cognitive and language skills at age 8 years.

At our two community labs (one in Northern California and one in the south bay of the San Francisco Bay Area), we conduct studies of parent-child interaction with both English- and Spanish-learning infants from families low in socioeconomic status (SES), who are growing up in very different circumstances from the high-SES English-learning participants at our Stanford campus lab. In both populations, significant disparities in language processing skill and vocabulary learning are already evident between higher- and lower-SES children by 18 months of age. And by the age of 24 months, there is a six-month gap between SES groups in processing skills that are known to be critical to language development.

LWL Stimuli: Choosing pictures for the target objects in the LWL task is complex. In addition to the requirement that the objects be easily identifiable from the picture, the choice of pictures depends partly on features of the individual picture of each object (size, position, angle, background, etc.) and partly on the relation between the two pictures in each pair. Each object is paired with another object - serving both as the target on 4 trials and the distracter on 4 trials. These paired pictures have to be balanced in visual salience. Otherwise the child's attention will be drawn more to one picture than the other just because it is more interesting - and not because of the child's language knowledge. To maintain the child's interest, we provide variation by using different pictures of each object, rather than showing the same picture of the object every time. Finally, we present each object during a "familiarization phase" that was added to the task with the goal of allowing the children to become familiar with the task before the testing started.

We selected eight target words to correspond to the images (see examples below). American 2-year-olds typically know the word "zebra" because of their experience with children's picture books. Their mothers use this word a lot in play and book reading, although most children have never seen a real zebra. In contrast, they see walls, floors, and trees every day, yet many 24-month-olds do not yet speak these words because their mothers don't talk to them about these particular objects. So the important question became “What objects do mothers in rural Senegal talk about with their young children?” These are the words their children are most likely to speak. For example, they may see goats and horses every day, but rarely hear them named. When choosing the words, and pairing them together in the LWL task, we took into consideration the following additional factors:
• Whether the Wolof name of the object is the same in all villages, or whether there are dialect variations in different villages. If the object can be named with different words, it is not a good choice (a concern with "telefon/portable").
• Whether the names of the objects in each pair begin with the same sound (like "cup" and "cat" in English). If so, then this is not a good pairing for the LWL task.
• Some objects are seasonal, e.g. mangos, so they will be differentially familiar to young children born in different seasons.
• In order to use a word such as “biscuits,” all the children needed to be familiar with the same brand of biscuit which was shown as a visual stimulus.

Our initial selection of target words was based on several sources of information, including a vocabulary list adapted for use in rural areas of Ghana, as well as early pilot data from Senegal. The final selection was made based on the combination of criteria above for the image and word selection, as well as the success of one of the STEP team members in capturing the images on camera.

**LWL Audio Recording:** In the LWL task, we recorded two sentences for the child to hear during each trial. The first sentence was the stimulus sentence with the target word, such as “Where’s the doggy?”, and the second sentence was an “attention getter”, such as “Do you like it?” We obtain the most important data immediately before, during, and immediately after the first sentence is played. The second sentence was used to keep the child engaged in the task. Trials in which the child was inattentive for the critical word or phrase of the stimulus sentence were prescreened out. In the above English example,
the noun, “doggy” is the only critical information in the sentence that the child needs to attend to. However, in Wolof, the definite article follows the noun and varies depending on the noun. “Where is the chicken” translates to “Ana ganaar gi?”, but “Where is the sheep?” translates to “Ana xar mi?”, where “gi” and “mi” are the definite articles. After discussions with our local team, we decided that it was not necessary for the child to hear the definite article in order to identify the correct image.

**LWL Piloting:** The process of adapting our lab-based version of the LWL procedure to a mobile version suitable for testing children in rural Senegal began during the summer of 2012. The procedure was originally programmed on a Macbook laptop computer, and LWL data collected from 4 Mandinka-speaking children and their mothers. The children were seated on their caregiver’s lap facing the computer. The laptop was placed on a table inside a cabana with obscure siding that still allowed light. Eye movements were recorded by a video camera attached to the top of the computer monitor. Although there was no time to select culturally appropriate stimulus pictures for the pilot, the speech stimuli were Mandinka words familiar to young children. The important result of this pilot project was that the Senegalese toddlers were intensely engaged by the images they saw in the LWL procedure, and the resulting data on the time course of word recognition were of high quality.

**LWL Coding:** Videotapes of children’s gaze were pre-screened trial-by-trial, and then coded frame-by-frame in the Dakar office, yielding high-resolution records of eye movements for each 33 ms interval. Trials were classified as either target-initial or distracter-initial depending on which picture the child was fixating at the onset of the target noun.

The LWL procedure yielded two measures of real-time speech processing: accuracy and reaction time (RT). We calculated accuracy as the mean proportion of looking to the named picture on both target- and distracter-initial trials, averaged over a 1500 msec window. The window started at 300 ms after the beginning of the target word being spoken (also referred to as noun onset), the minimum time it would take the child to respond meaningfully to a sound, and ended at 1800 ms after noun onset.

Mean reaction time (RT) was calculated as the child’s latency to shift from looking at the distracter to the target picture. We only included trials on which the child was looking at the distracter picture at the onset of the target word and shifted to look at the target picture within 300-1800 ms of noun onset. Trials on which the child shifted either within the first 300 ms or later than 1800 ms from target word onset were excluded, since these early and late shifts were less likely to be in response to the stimulus sentence.

To assess reliability, a highly experienced ‘gold-standard’ coder independently recoded 10% of the trials. Inter-observer agreement was computed in two ways: (1) percentage of all trials for which the accuracy estimate for the coders agreed within 0.05, and (2) percentage of all trials for which the reaction time estimates for the coders agreed within 2 frames or 67 msec. These statistics were computed from the first half of the coding period from baseline, and showed that the coders had very high reliability. Specifically, the agreement for accuracy was 98%, for distracter-initial trial RT was 99%, and for target-initial trial RT was 97%.
In addition to the two summary statistics, graphical representations demonstrate the time course of looking to the target picture as the stimulus sentence unfolds during the trials. Specifically, the mean proportion of trials that children focus on the target picture are plotted (at 33 ms intervals), averaged over all children. Looking to the target picture is expected to be near chance (i.e., 50%) when the images first appear (picture onset) and before there is sufficient audio information to enable the child to identify the target image. After noun onset, we expect the mean proportion of correct looking to target to begin to rise and then to flatten as children focus on the correct image. After the attention getter sentence, the children are expected to lose interest in the target image, and the proportion will start to drop.

Parents’ report of children’s language proficiency To assess how children’s communication skills developed over time in the RPP and comparison groups, two caregiver report measures were used: a language-milestone checklist for infants and young children of increasingly more advanced language skills and an inventory of vocabulary words that young Wolof-speaking children in Kaolack can typically understand and say.

Language milestones: Children’s language development begins early in infancy, long before the emergence of the first word. The language milestone checklist was developed to assess the communication skills of even the youngest children in our study. Specifically, we adapted and translated an English-version checklist available with the LENA™ device into Wolof for use in the Senegalese context. The checklist consists of a series of questions (items) regarding increasingly more advanced language skills in infants and young children. We also consulted other milestone checklists for additional items to include. Mothers were asked to report whether or not their child was evidencing each of the skills, and their response was scored with a simple 0 (no) or 1 (yes). All children were started with the first item and a stopping rule of 6 no’s in a row was applied. Language milestones achieved are reported as the raw total score of the items administered. Items not administered after the stopping rule were given a score of 0.

Caregivers for 20 children were administered the Milestone checklist as part of a small pilot study in 3 Wolof-speaking villages in rural Senegal. The children ranged in age from 2.5 months to 22 months of age. Only 2 items didn’t perform as expected based on Item Response Modeling (IRM) analysis, and were either re-worded or dropped. Specifically, one of these items, “Does your child bring toys or objects to his/her mouth?” was identified during piloting to be a problem and was dropped. Respondents thought that we wanted to know whether their children put “dirty” things in their mouths, so they said no, their child does not do that. The ordering of items by difficulty was reasonable and was not changed. Language skill was strongly correlated with age in this small sample (rho = 0.91).

The final language milestone checklist at baseline consisted of a series of 38 items. Since all children were one year older at follow-up, the earliest milestones were dropped (e.g., “Child babbles or turns head to the sound of their mother’s voice”) and added more advanced items (e.g., “Child can describe two things about a named object”), for a total of 40 items at both follow-up phases. The person separation reliabilities were excellent in both 2013 and 2014 (.87 at phase 1 and .94 at phase 2),
indicating that the instrument was sensitive enough to distinguish between high and low performers on the milestones checklist.

**Expressive Vocabulary:** The MacArthur Communicative Development Inventory (CDI) is the most widely used instrument for assessing vocabulary development over the first three years of life (Fenson et al., 2007). The development of a Wolof CDI for the baseline study was a multi-step process for which we obtained the approval of the CDI Advisory board (Weber & Marchman, 2013). In the first step, a pilot version of the Wolof CDI was developed using an existing West African version of the CDI as a starting point (Prado, 2011). The West African CDI was developed and adapted in 2010 for use with the languages of Krobo, Ewe and Twi, which are spoken primarily in southern regions of Ghana and Malawi, and belong to a different language family from Wolof. However, the socio-cultural and economic conditions were thought to be comparable to Senegal, making the West African CDI an excellent starting point for the Wolof version.

In consultation with Wolof-speaking members of the Tostan staff, we examined each item for cultural and linguistic relevance to Wolof, and substituted those items that were not appropriate to our study population. Items were avoided that are homonyms and which may be particularly variable across dialects. We also consulted other versions of the CDIs and added items as necessary. This step culminated in the development of a pilot version with 120-130 potential items.

In the second step, data were collected using the pilot form from 30 Wolof-speaking mothers of young children following an interview format. Mothers were told that we would like to know whether their child “understands and says” each of the words. It was explained that interviewers were not interested in words that the child can only imitate and that other words that may differ as a function of dialect or family are also acceptable. When mothers reported a positive response, mothers were then asked to provide examples of specific times when their child used that word. These probes served to make sure that the mothers understood our instructions. Mothers were asked to provide examples of additional words that were not on the list but that they have heard their child say. As the result of this piloting, a second version of the form was developed. We removed items that no child spoke, and added some items that were provided by the mothers. Following the CDI published guidelines we developed a short form of 105 Wolof words likely to be familiar to children 20-30 months, and that reflected a distribution of approximately 20% easy, 60% middle, 20% hard items. For the phase 2 follow-up, the list was modified by replacing the easiest words (e.g., *milk*) with a set of harder words (e.g., *before*) for a total of 110 words.

Expressive vocabulary is reported as the raw total of the CDI word scores, where each word was scored as 0 for “no” and 1 for “yes.” As with the milestones checklist, the person-separation reliabilities were excellent for both forms and for the younger form in 2013 and 2014 (exceeding .91 in all cases), indicating that the instrument was able to distinguish between high and low performers on the CDI. Cronbach’s alpha was also excellent at .97 to .99.
Other Measures

Communities that received the program were likely to be different from the comparison villages due to the sampling strategy simply by chance. Given the potential confounding effect of such differences, extensive data were collected to control for the differences statistically. Questionnaires were developed to gather information related to village infrastructure, adult education, household socio-economic status, and child health.

In order to improve the reliability of the survey data, all questionnaires were translated and transcribed into Wolof - a significant change from common practice in survey research in Senegal since most Senegalese adults, include interviewers, are not literate in their native language. Questionnaires are typically written in French and then translated “online” into Wolof as the survey is being administered to local participants. This practice can result in substantial variability due to differences in interviewers’ wording of the questions, as well as differences in their translations over time. However, the translation process was time-intensive and iterative. The questionnaires were initially translated by Tostan translators, but went through many corrections and changes during training and pilot testing based on feedback from the STEP team. We also worked with Senegalese mothers and other Wolof language experts to verify that our assessment tools were culturally and linguistically appropriate for the dialect used in Kaolack. These survey tools are available upon request in both French and Wolof.

Village: For the village questionnaire, a group of community leaders was asked to give basic demographic information on the village, including population size, and number of households, midwives and marabouts. They provided information on access to services, such as clinics and schools, availability of electricity and running water, and participation in other programs besides the RPP program. In addition, they were asked about the occurrence of major economic shocks that affected the whole village, such as flooding or loss of crops.

Household: The household questionnaire consisted of a set of standard questions that have been used previously for the Demographic Health Surveys (DHS) and by other organizations in Senegal. The main purpose of the household questionnaire was to understand (and control for) possible differences between the communities and families that might explain differences in outcomes across study arms. Specifically, we obtained the household size, number of children under 5 years of age living in the household, the median adult education (for household members > 9 years of age), participation of members in the Tostan CEP program, and the ethnicity and marital status of the household members, including whether the head of household was in a polygynous marriage.

In addition, information was obtained on multiple wealth-related variables (i.e. ownership of durable goods and livestock assets, and dwelling conditions such as electricity, running water, composition of floor, walls, and roof). A household wealth index (mean 0, SD 2.5) was generated from these variables using principal component analysis to aggregate them into a single measure. The first principal component was retained as it captures the most common variation among the variables. The reliability of the index to yield household rankings in wealth has long been established for the DHS surveys and validated with rankings using household expenditures (Vyas & Kumaranayake, 2006).
Primary Caregiver Identification: The “responsable” questionnaire was administered to the person who was most well informed (usually makes decisions) about the child’s health and care. The purpose of this questionnaire was to understand the child care-giving practices, as well as to obtain some basic indicators of the child’s health status that may have influenced the child’s performance on the assessments. Importantly, the questionnaire was used to identify the primary caregiver - defined as the individual who spent the most daylight hours per day caring for the child; in other words, the person who had the most opportunity to talk with the child and influence his or her language development. In most cases, this was the mother, but since many young mothers in rural Senegal have other duties during the day, such as agricultural work, household tasks and/or income-generating activities, care may have been provided by an older sibling or adult relative.

Caregiver Characteristics: Caregiver education was assessed by asking the caregiver about her highest achieved level of formal education in French or French-Arabic schools, and was then categorized into one of three categories: no education, some primary education, or secondary education and above. Caregivers were also asked about prior participation in the Tostan Community Empowerment Program (CEP) and in any parental education programs.

Reproductive history was assessed by asking the caregiver about the number and sex of her surviving children, the number and sex of children who were live births but had since died, and whether she was pregnant at the time of the interview.

Caregiver time use was assessed through a series of questions asking the caregiver to identify her principle and secondary activity during various times of the day, as well as whether the child was with her during each time period. Days were divided into eight time periods and activities were divided into the following categories, each of which had several sub-categories: household tasks, child-care tasks, income-generating tasks, and other tasks (e.g., sleeping, religious activities, going to school etc.). Caregivers were also asked whether they looked after any other children aged five or younger.

Self-reported levels of depression for primary caregivers was adapted from a 10-question short form of the Center for Epidemiologic Studies Depression (CES-D) scale (Grzywacz, 2006; Radloff, 1977). The interviewer recited an introductory statement that taking care of children requires a lot of time and energy and that those who take care of children sometimes feel tired or overwhelmed. The primary caregiver was then asked to identify at what level she had felt any of the emotions or symptoms read by the interviewer over the past week (e.g., feeling alone, happiness, sadness, difficulty sleeping, etc.). Response options were Likert-style 4-point options from never/rarely to most of the time.

Child Health & Nutrition: Nutritional status was obtained by measuring children’s height and weight, converted into age-adjusted z-scores using WHO growth standards and software (World Health Organization, 2011). In addition, the “responsable” was asked a series of questions related to feeding practices, which included: the child’s breastfeeding history; the age at which the child was introduced to solid and semi-solid foods; consumption of various solid and semi-solid foods over the past 24 hours; and whether or not they had been given any iron supplements over the past week, or Vitamin A supplements or medication against worms over the past six months.
Interviewers asked the “responsable” of each child about recent illness, specifically if the child had been sick with a cough, fever, or diarrhea over the past 14 days. Those who sought medical advice concerning the child over that time period were asked to whom they went (e.g., hospital, pharmacist, traditional healer etc.) and, if a child had been sick, interviewers asked about the extra amount of time family members spent to care for the sick child. Vaccination history was determined by interviewers through consultation of the child’s vaccination card and, when necessary, through mother report.

Section 4 – Analyses

Raw Data Distributions

For many of the outcome measures, we present raw data distributions for each program group and each survey period graphically using kernel density plots (a non-parametric, effective alternative to a histogram). We overlay vertical lines to indicate the location of the mean for each sub-sample. These plots provide a visual representation of differences that may exist between the program and comparison villages, as well as before and after the program implementation. We refer to the difference between the post-program outcomes in program vs. comparison groups as the “unadjusted effect estimate” as it does not control for confounding by the lagged pre-program outcome or other possible confounders. It is important to note that the error of the estimates of the means are not shown; thus these plots do not tell us if the differences are statistically significant. The contour line style and color are the same for all distribution plots: short-dash for phase 1, solid for phase 2, and long dash for phase 3; blue for the RPP group and red for the comparison group.

Regression Estimates of Effect

Our target parameter of interest for the evaluation is the average treatment effect (ATE) between the RPP and comparison groups. We define the ATE as the difference in mean post-program outcomes for the two groups. In the following section, the mean outcome for a group is denoted as: \( E(Y_{i}(p) \mid p, T, W, Y_{i}(p=1)) \), where \( Y_{i}(p) \) is the individual, \( i \), outcome at phase \( p=1, 2 \) or \( 3 \); \( T \) is the number of years of exposure to the program or treatment, \( T=0, 1, \) or \( 2 \); and \( W \) represents potential confounders. Despite efforts to minimize differences across groups, communities that received the program were found to be systematically different from the comparison villages at baseline for a number of key factors (e.g., child nutritional status). Given these potential confounding effects, significant associations found with the raw unadjusted data may be a result of bias. We address confounding by including these factors in parametric linear regressions. Importantly, the phase 1 outcome measures, \( Y_{i}(p=1) \), also known as lagged outcomes, are always included in the regression of the corresponding post-RPP outcome measure.¹ For example, in the regression on quantity of caregiver talk post-RPP, we control for the

¹ For all regressions, we included the residuals on baseline outcomes adjusted for age. Phase 3 sampling was a function of age at baseline (only the younger cohort was followed), which was predictive of baseline scores. Inclusion of both an indicator for two years of program participation and baseline outcome in the same model would bias the estimate of the impact of program participation.
quantity of caregiver talk pre-RPP along with the other key covariates, such as child and caregiver demographics.

In this report, we present three estimates, $\Psi$, for the caregiver and child outcomes. The first equation below for $\Psi^I$ represents the expected difference in mean outcome, $Y$, at phase $p=2$ for one year of exposure to the RPP ($T=1$) and mean outcome at $p=2$ for no exposure for the controls ($T=0$), adjusting for covariates.

$$
(1) \Psi^I(P_0) = E_{W,Y(p=1)} \left( \frac{E(Y_i(p=2)|T=1, W_i, Y_i(p=1)) - E(Y_i(p=2)|T=0, W_i, Y_i(p=1))}{-} \right)
$$

The next equation represents the expected mean difference in outcome at phase 3 for two years of exposure to the RPP ($T=2$) and outcome at phase 2 with no exposure for the controls ($T=0$).

$$
(2) \Psi^{II}(P_0) = E_{W,Y(p=1)} \left( \frac{E(Y_i(p=3)|T=2, W_i, Y_i(p=1)) - E(Y_i(p=2)|T=0, W_i, Y_i(p=1))}{-} \right)
$$

The third equation represents the expected mean difference in outcomes at phase 3 for two years of exposure ($T=2$) and one year of exposure ($T=1$) to the RPP.

$$
(3) \Psi^{III}(P_0) = E_{W,Y(p=1)} \left( \frac{E(Y_i(p=3)|T=2, W_i, Y_i(p=1)) - E(Y_i(p=2)|T=1, W_i, Y_i(p=1))}{-} \right)
$$

To obtain the program effect estimates, we imposed the following generalized form for the linear regression estimator:

$$
E(Y_i(p)|T, W_i, Y_i(p=1)) = \beta_0 + \beta_1 T + \beta_2 W_i + \beta_3 Y_i(p=1) \quad i = 1, \ldots, I \quad p = 2,3
$$

The estimated ATE included in the results (Tables 3 & 4) are obtained from the regression coefficient, $\beta_1$, on treatment, $T$. Child age concurrent with the outcome measure was used in the regressions. In all cases, standard error of estimates were adjusted for clustering at the village level and for repeated measures at the individual level. Missing covariate information (i.e., confounder variables that were not pre or post outcome measures) were imputed.

We present the covariate-adjusted RPP effect estimates using the original metric (e.g., number of words, points) for caregiver outcomes (Table 3), outcomes administered to children of all ages at baseline (Top of Table 4), and outcomes administered only to the older cohort of children at baseline (i.e., vocabulary and LWL, bottom of Table 4).

Section 5 – Results
At baseline in 2013, we collected audio, visual, and quantitative data for 506 caregiver-child pairs in 423 households, which typically consisted of large extended families living together in small buildings set around a central compound with a communal kitchen. Children were 4 to 31 months of age.
One year after baseline, when the children were 16-43 months, 469 (92.7%) of these caregiver-child pairs were tested again (29 had moved away, 7 declined, one child was deceased). For the analyses, we excluded one child with a developmental delay and 25 children whose primary caregiver had changed between the two study periods (n=443 remaining). Most primary caregivers (84.7%) had no formal education in French or French-Arabic schools, 11.5% had some primary education, and less than 4% had any secondary or higher education. Because 66% of heads of household were in polygynous marriages, households were large, averaging 16.1 members and 4.1 children under the age of 5 years. Less than half of the villages had a functional health center and only three had electricity (Table 1).

At phase 3, the team located and tested 91 of the 94 caregiver-child pairs targeted in the RPP villages; three had moved away (Dakar, Gambia, and the Casamance). One additional child had a new caregiver for more than 6 months and was excluded from the analyses. All but 3 of these caregivers had participated in the new RPP activities in 2014.

The sample size used in the regression analyses varied by outcome based on the availability of data at each phase, from 177 (e.g., for LENA™ measures) to 428 (e.g., for play session measures) caregiver-child pairs included in the effect estimates.

**Baseline Comparison of RPP and Control Groups**

The goal of this first quantitative analysis was to test whether caregiver-child pairs in RPP and comparison villages were comparable across potential confounders, indicating that the villages were adequately matched for the evaluation. In general, the distribution of caregiver outcomes (e.g., behavior, knowledge, and attitude) are very similar for the RPP and comparison groups at the pre-program baseline period. These findings were re-assuring as they suggested that the RPP and comparison caregivers were similar along these dimensions despite the fact that the program was not assigned randomly to the villages.

Balance between the two groups was tested statistically with either a t-test (continuous variables) or chi-square test (dichotomous or categorical variables), adjusting the standard error to account for the correlated nature of our data at the village level. No significant differences were found between groups for child age, caregiver education and relationship to the child, or for household size and wealth. However, children in the program villages were more likely to be male (59.8 vs. 49.3%) and to have a primary caregiver who was slightly older (29 vs. 27 years) than were children in the comparison villages (Table 1). The number of caregivers who had completed Tostan’s foundational Community Empowerment Program was also higher in program villages (54.9 vs. 30.6%) because the CEP had been held in many (but not all) of the RPP villages more recently than in the comparison villages. There were no statistically significant differences between the two groups for any of the key caregiver behavior and child language measures assessed at baseline (Table 2, columns 2 and 3 for 2013). However, caregiver attitude to discipline score was lower by ~1 point in the RPP group (RPP caregivers were less in favor of the use of corporal punishment), and this difference was statistically significant (p-value=.037). This difference may be due to the higher participation rate of RPP caregivers in Tostan’s CEP program, which has a focus on human rights.
Table 1: Baseline demographics of child, caregiver, household, and village by program status

<table>
<thead>
<tr>
<th></th>
<th>Comparison</th>
<th>RPP</th>
<th>All</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Number of dyads</strong></td>
<td>219</td>
<td>224</td>
<td>443</td>
</tr>
<tr>
<td><strong>Child Characteristics</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Age in months</td>
<td>17.2 (7.5, 4-30)</td>
<td>18.6 (7.3, 4-31)</td>
<td>17.9 (7.4, 4-31)</td>
</tr>
<tr>
<td>Male</td>
<td>49.3% (108)</td>
<td>59.8% (134)</td>
<td>54.6% (242)*</td>
</tr>
<tr>
<td>Birth size estimate</td>
<td>2.9 (1.1, 1-5)</td>
<td>2.8 (1.3, 1-5)</td>
<td>2.9 (1.2, 1-5)</td>
</tr>
<tr>
<td>Length-for-age z-score</td>
<td>-1.1 (1.2, -4.7-2.8)</td>
<td>-1.4 (1.2, -4.9-2.9)</td>
<td>-1.3 (1.2, -4.9-2.9)</td>
</tr>
<tr>
<td><strong>Caregiver Characteristics</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Relationship to child</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Mother</td>
<td>95.0% (208)</td>
<td>92.9% (208)</td>
<td>93.9% (416)</td>
</tr>
<tr>
<td>Step-mother</td>
<td>1.8% (4)</td>
<td>0.4% (1)</td>
<td>1.1% (5)</td>
</tr>
<tr>
<td>Grandmother</td>
<td>1.8% (4)</td>
<td>4.5% (10)</td>
<td>3.2% (14)</td>
</tr>
<tr>
<td>Sibling</td>
<td>0.5% (1)</td>
<td>0.0% (0)</td>
<td>0.2% (1)</td>
</tr>
<tr>
<td>Aunt</td>
<td>0.9% (2)</td>
<td>2.2% (5)</td>
<td>1.6% (7)</td>
</tr>
<tr>
<td>Age in years</td>
<td>26.9 (7.5, 15-60)</td>
<td>29 (8.8, 16-67)</td>
<td>28 (8.2, 15-67)*</td>
</tr>
<tr>
<td>Years of Koranic schooling</td>
<td>4 (2.7, 0-10)</td>
<td>4.1 (2.8, 0-10)</td>
<td>4.1 (2.7, 0-10)</td>
</tr>
<tr>
<td>Formal education in French or Arabic schools</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>None or preschool</td>
<td>82.2% (180)</td>
<td>87.1% (195)</td>
<td>84.7% (375)</td>
</tr>
<tr>
<td>Some primary</td>
<td>13.2% (29)</td>
<td>9.8% (22)</td>
<td>11.5% (51)</td>
</tr>
<tr>
<td>Some secondary or higher</td>
<td>4.6% (10)</td>
<td>3.1% (7)</td>
<td>3.8% (17)</td>
</tr>
<tr>
<td>Does not know letters in French or Wolof</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Completed Tostan CEP</td>
<td>30.6% (67)</td>
<td>54.9% (123)</td>
<td>42.9% (190)*</td>
</tr>
<tr>
<td>Number of children under 5 years she cares for</td>
<td>2.8 (2.1, 1-13)</td>
<td>3 (2, 1-11)</td>
<td>2.9 (2, 1-13)</td>
</tr>
<tr>
<td>Has lost a child</td>
<td>34.7% (76)</td>
<td>40.6% (91)</td>
<td>37.7% (167)</td>
</tr>
<tr>
<td>Reported depressive symptoms score (0-30 possible)</td>
<td>7.5 (5.5, 0-27)</td>
<td>7 (5.2, 0-27)</td>
<td>7.2 (5.3, 0-27)</td>
</tr>
<tr>
<td><strong>Number of Households</strong></td>
<td>190</td>
<td>193</td>
<td>383</td>
</tr>
<tr>
<td>Household Characteristics</td>
<td>Comparison</td>
<td>RPP</td>
<td>All</td>
</tr>
<tr>
<td>--------------------------</td>
<td>------------</td>
<td>-----</td>
<td>-----</td>
</tr>
<tr>
<td>Number of members</td>
<td>17.1 (7, 5-42)</td>
<td>15.2 (6.3, 4-41)</td>
<td>16.1 (6.7, 4-42)</td>
</tr>
<tr>
<td>Number of children &lt;= 5 years</td>
<td>4.4 (2.3, 1-13)</td>
<td>3.8 (2.2, 1-11)</td>
<td>4.1 (2.3, 1-13)</td>
</tr>
<tr>
<td>Mean age (years) of household members</td>
<td>19.5 (3.4, 10.7-30.2)</td>
<td>19.9 (3.7, 11.7-30.2)</td>
<td>19.7 (3.5, 10.7-30.2)</td>
</tr>
<tr>
<td>Proportion with median adult education = none</td>
<td>95.3% (181)</td>
<td>88.1% (170)</td>
<td>91.6% (351)</td>
</tr>
<tr>
<td>Number of adults who participated in Tostan’s CEP</td>
<td>1.7 (1.5, 0-7)</td>
<td>2 (1.8, 0-15)</td>
<td>1.9 (1.7, 0-15)</td>
</tr>
<tr>
<td>Proportion with polygynous head of household</td>
<td>71.6% (136)</td>
<td>60.1% (116)</td>
<td>65.8% (252)</td>
</tr>
<tr>
<td>Household wealth indicator</td>
<td>0.5 (2.2, -4.5-5.1)</td>
<td>-0.6 (2.7, -5.5-5.2)</td>
<td>0 (2.5, -5.5-5.2)</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Number of Villages</th>
<th>12</th>
<th>12</th>
<th>24</th>
</tr>
</thead>
<tbody>
<tr>
<td>Village Characteristics</td>
<td>% (N) or Mean (SD, range)</td>
<td>% (N) or Mean (SD, range)</td>
<td>% (N) or Mean (SD, range)</td>
</tr>
<tr>
<td>Estimated population size</td>
<td>1103 (570, 460-2322)</td>
<td>842 (518, 200-1878)</td>
<td>973 (549, 200-2322)</td>
</tr>
<tr>
<td>Estimated number of households</td>
<td>244 (276, 45-999)</td>
<td>140 (65, 52-250)</td>
<td>192 (203, 45-999)</td>
</tr>
<tr>
<td>Number of midwives</td>
<td>1.6 (1, 0-3)</td>
<td>1.2 (0.8, 0-3)</td>
<td>1.4 (0.9, 0-3)</td>
</tr>
<tr>
<td>Number of marabouts</td>
<td>3.9 (3, 1-10)</td>
<td>2.8 (2.2, 1-8)</td>
<td>3.4 (2.6, 1-10)</td>
</tr>
<tr>
<td>Has a functional health facility</td>
<td>41.7% (5)</td>
<td>50.0% (6)</td>
<td>45.8% (11)</td>
</tr>
<tr>
<td>Presence of a primary or secondary school</td>
<td>83.3% (10)</td>
<td>83.3% (10)</td>
<td>83.3% (20)</td>
</tr>
<tr>
<td>Has a day care facility</td>
<td>16.7% (2)</td>
<td>8.3% (1)</td>
<td>12.5% (3)</td>
</tr>
<tr>
<td>Has electricity service to the village</td>
<td>8.3% (1)</td>
<td>16.7% (2)</td>
<td>12.5% (3)</td>
</tr>
<tr>
<td>Time to travel to market</td>
<td>2 (0.6, 1-3)</td>
<td>1.6 (1, 0-3)</td>
<td>1.8 (0.8, 0-3)</td>
</tr>
<tr>
<td>Time to obtain drinking water</td>
<td>1.2 (1.8, 0-4)</td>
<td>0.5 (0.9, 0-3)</td>
<td>0.8 (1.4, 0-4)</td>
</tr>
<tr>
<td>Number of other programs in village</td>
<td>0.8 (0.8, 0-2)</td>
<td>1.1 (0.7, 0-2)</td>
<td>1 (0.8, 0-2)</td>
</tr>
<tr>
<td>Suffered from flooding in 2013</td>
<td>75.0% (9)</td>
<td>91.7% (11)</td>
<td>83.3% (20)</td>
</tr>
</tbody>
</table>

* Indicates a statistically significant difference (p-value < 0.05) between the RPP and comparison groups at baseline, using a 2-sided t-test (continuous variables) or a chi-squared test (binary or categorical variables), adjusted for clustering at the village level.

a Persons over 15 years of age.
Table 2: Unadjusted, descriptive statistics for the outcome measures by program status and survey year

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Caregiver Outcomes</td>
<td>Mean (SD, range)</td>
<td>Mean (SD, range)</td>
<td>Mean (SD, range)</td>
<td>Mean (SD, range)</td>
<td>Mean (SD, range)</td>
</tr>
<tr>
<td>Caregiver words to child</td>
<td>226 (131, 1-629)</td>
<td>228 (151, 0-701)</td>
<td>224 (114, 2-518)</td>
<td>405 (158, 34-838)</td>
<td>397 (140, 55-686)</td>
</tr>
<tr>
<td>Caregiver MLU</td>
<td>2.54 (0.69, 1-5.21)</td>
<td>2.55 (0.93, 1-11.06)</td>
<td>2.72 (0.56, 1.46-4.32)</td>
<td>3.41 (0.76, 1.26-6.03)</td>
<td>3.43 (0.75, 2-5.51)</td>
</tr>
<tr>
<td>Female adult words/hr</td>
<td>517 (269, 136-1240)</td>
<td>533 (304, 9-2243)</td>
<td>365 (223, 57-1593)</td>
<td>362 (218, 31-1287)</td>
<td>317 (182, 4-686)</td>
</tr>
<tr>
<td>Knowledge of child development</td>
<td>23.7 (4.8, 12-40)</td>
<td>23.1 (4.4, 12-41)</td>
<td>27.5 (4.8, 13-39)</td>
<td>36.3 (4.2, 23-45)</td>
<td>35.8 (4.1, 27-43)</td>
</tr>
<tr>
<td>Attitude to discipline</td>
<td>19.2 (4.2, 6-24)</td>
<td>18 (4.6, 3-24)</td>
<td>16.8 (5.6, 0-24)</td>
<td>10.8 (6.8, 0-24)</td>
<td>11.5 (6.7, 0-24)</td>
</tr>
<tr>
<td>Child Outcomes</td>
<td>Mean (SD, range)</td>
<td>Mean (SD, range)</td>
<td>Mean (SD, range)</td>
<td>Mean (SD, range)</td>
<td>Mean (SD, range)</td>
</tr>
<tr>
<td>Child utterances</td>
<td>24.7 (31.9, 0-150)</td>
<td>30.3 (31.7, 0-127)</td>
<td>51.7 (33.8, 0-149)</td>
<td>65.1 (33.1, 0-158)</td>
<td>67.1 (31.3, 0-117)</td>
</tr>
<tr>
<td>Language milestones</td>
<td>16.2 (6, 4-38)</td>
<td>17.5 (5.9, 5-34)</td>
<td>18.3 (6.9, 5-32)</td>
<td>21.8 (7.6, 5-37)</td>
<td>27.6 (5.7, 13-40)</td>
</tr>
<tr>
<td>Expressive vocabulary</td>
<td>44.2 (29.2, 2-103)</td>
<td>53.7 (26.4, 5-101)</td>
<td>69.3 (19.1, 1-105)</td>
<td>74.5 (20, 0-103)</td>
<td>79.5 (17.6, 28-107)</td>
</tr>
<tr>
<td>Child vocalizations/hr</td>
<td>219 (91, 60-439)</td>
<td>235 (90, 43-489)</td>
<td>287 (99, 92-512)</td>
<td>324 (102, 116-602)</td>
<td>289 (123, 22-476)</td>
</tr>
<tr>
<td>Conversation turns/hr</td>
<td>54.3 (22.3, 22.1-110)</td>
<td>58.1 (21.8, 7.5-130)</td>
<td>52.4 (20.7, 21.8-113)</td>
<td>60.8 (23.5, 14.2-129)</td>
<td>51 (26, 1.1-99)</td>
</tr>
<tr>
<td>LWL: accuracy</td>
<td>0.55 (0.13, 0.21-1)</td>
<td>0.57 (0.1, 0.21-0.92)</td>
<td>0.56 (0.1, 0.12-0.8)</td>
<td>0.57 (0.09, 0.26-0.8)</td>
<td>0.59 (0.11, 0.25-0.83)</td>
</tr>
<tr>
<td>LWL: reaction time</td>
<td>892 (214, 367-1600)</td>
<td>855 (210, 300-1567)</td>
<td>890 (189, 300-1533)</td>
<td>861 (196, 400-1489)</td>
<td>862 (216, 489-1633)</td>
</tr>
</tbody>
</table>

*aRaw, unadjusted, numbers are reported here for descriptive purposes. Participant counts (N) vary by outcome, program status and survey year.
Changes in caregivers’ parenting skill

Caregiver’s speech to child: 1- and 2-year program effect.  At baseline, caregivers in all 24 villages spoke on average about the same amount to their child during the 5 min structured play session, as shown with the closely matched distributions in Figure 2 (dotted lines). One year after the baseline study, at Phase 2 (solid lines), the entire distribution of talk by caregivers in program villages had shifted upwards. In contrast, caregivers in comparison villages showed no substantial change in amount of child-directed talk. Two years after the baseline study (long dashed lines), at Phase 3, the amount of talk in program villages remained high, but did not increase beyond the first year gains.

Caregivers in the RPP group also used more complex language at follow-up (Figure 3), increasing their mean length of utterance (MLU) after one year, with no change in the comparison group. As with the amount of talk, the gain in MLU was maintained in the program group after another year.

The substantial first year program effect remains nearly the same when adjusted statistically for

Figures 2 and 3.  The number of caregiver words and mean length of utterance, respectively, in the 5 min structured play session, by Treatment (program) and Control (comparison) groups, at baseline (dotted lines), and one (solid lines) and two years (long dashed lines) after program implementation. Vertical lines represent the mean for each sub-group.
key factors that could bias the results, including amount of caregiver speech at baseline (Table 3, 1-year effect). The estimated gains after 2 years of participation in the RPP were smaller than the one-year effects as compared to the comparison group, although the effect estimates were still significant for caregiver words to the child and mean length of utterance in the play session (Table 3, 2-year effect estimates).

Although small, the decrease in effect size from phase 2 to phase 3 in the RPP groups was borderline statistically significant for caregiver words to the child in the play session (Table 3, additional year estimate).

Table 3: Adjusted\(^a\) RPP program effect estimates for key caregiver outcome measures

<table>
<thead>
<tr>
<th>Caregiver Outcomes(^c)</th>
<th>1-year Effect estimate, $\Psi_I$</th>
<th>2-year Effect estimate, $\Psi_II$</th>
<th>Additional year estimate(^b), $\Psi_{III}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Play session: Caregiver words to child</td>
<td>181 (13)**</td>
<td>151 (20)**</td>
<td>-30 (18)*</td>
</tr>
<tr>
<td>Play session: Caregiver MLU</td>
<td>0.72 (0.12)**</td>
<td>0.55 (0.13)**</td>
<td>-0.16 (0.13)</td>
</tr>
<tr>
<td>All-day audio: Female adult words/hr</td>
<td>-34.9 (59.2)</td>
<td>-29.2 (72.0)</td>
<td>5.7 (46.6)</td>
</tr>
<tr>
<td>Knowledge of child development</td>
<td>7.3 (0.5)**</td>
<td>5.9 (0.6)**</td>
<td>-1.4 (0.6)**</td>
</tr>
<tr>
<td>Attitude to discipline</td>
<td>-4.8 (0.7)**</td>
<td>-3.1 (1.1)**</td>
<td>1.8 (0.9)**</td>
</tr>
</tbody>
</table>

\(^a\) Covariates are age-adjusted lagged dependent variables plus characteristics of the child (age, change in age, age squared, gender, length-for-age, birth size), caregiver (age, relation to child, education, literacy, attended earlier Tostan CEP, children < 5 years in her care, lost a child, depressive symptoms), household (polygynous head of household, members who attended earlier Tostan CEP, mean age, median years of education, children < 5 years, crowding, wealth), and village (population, households, midwives, and marabouts; health clinic, primary school, day care time to market and to get water, electricity, operational NGOs, flooding in 2013).

\(^b\) Difference in effect estimates between one and two years of exposure

\(^c\) Caregiver-child play session: caregiver words to child, Mean Length of Utterance (MLU); All-day recording using LENA\(^TM\): Female adult words to child; Caregiver report questionnaire data: Caregiver knowledge of child development; Caregiver attitude toward the use of corporal punishment.

* p-value <.1, ** p-value <.05, *** p-value<.01
Figure 4 shows that adult female-initiated speech from the all-day LENA™ recordings were also comparable in program and comparison villages at the Phase 1 baseline (overlapping dotted line distributions). However, the distribution and mean count per hour shifted downwards after one year in both groups (solid lines), dropping further after the second year in the program villages (long dashed lines). No statistically significant differences were found between the two groups at any phase (Tables 2 and 3).

Through further exploration, we found that change in amount of talk was associated with child age and amount of baseline talk, but not a function

**Figure 4.** The adult female-initiated word count per hour from the all-day LENA™ recordings by Treatment (program) and Control (comparison) groups, at baseline (dotted lines), and one (solid lines) and two years (long dashed lines) after program implementation.

**Figure 5:** Caregivers in the RPP group who talked more at baseline changed less than those in the RPP group who talked less at baseline (left). **Figure 6:** The amount of caregiver talk after one year in the RPP flattens out as amount of talk at baseline increases (right).
of age at start of program (i.e., cohort) or one vs. two years of participation. Caregivers of older children talked more on average. Also, caregivers in the RPP group who talked more at baseline also talked more at follow-up, but they changed less than those in the RPP group who talked less at baseline (see Figure 5). This may be a function of a ceiling effect, as suggested in Figure 6, where we show a flattening in the amount of caregiver talk at phase 2 as amount of talk at baseline increases.

**Caregiver knowledge & attitude toward discipline.** The distribution of caregiver knowledge of child development is normal and similar for the RPP and comparison groups at baseline (Figure 7). Although both groups of caregivers showed increased knowledge of child development when interviewed one year later, caregivers in RPP villages improved significantly more (Tables 2 and 3).

As with caregiver knowledge, Figure 8 shows that the distribution of discipline scores are similar for the RPP and comparison groups at baseline, but the distribution is skewed, with both groups strongly in favor of the use of physical punishment: 50% of caregivers scored 20 or more points out of a

![Diagram showing the distribution of discipline scores for caregivers in the RPP and comparison groups at baseline, indicating a skew towards higher scores for the use of physical punishment.](image)

Figures 7 and 8. Caregiver knowledge of child development (top) and attitude toward the use of physical punishment (bottom) by Treatment (program) and Control (comparison) groups, at baseline (dotted lines), and one (solid lines) and two years (long dashed lines) after program implementation.
maximum possible of 24 (or 83% of the maximum score). A higher discipline score equates to agreeing more strongly to a parent hitting their child for such situations as: disobeying an elder, fighting with others, refusing to do chores, etc. In order to achieve a score of 20 points, the caregiver might have responded “strongly agree” to the use of hitting in 4 of the situations and “agree” to the use of hitting in the other 4 situations (out of 8 total situations presented). At phase 2, the mean discipline scores decreased on average in both groups, but the distribution in the comparison group remained highly skewed with most respondents continuing to strongly favor the use of physical punishment. Although the mean discipline score for the RPP group shifted down significantly, the distribution of the scores in this group is nearly even across the full range of possible scores, with many mothers in the RPP group continuing to express favor toward the use of corporal punishment at home. Only about 10% of the caregivers who participated in the RPP program scored a zero on the discipline scale (they responded strongly disagree to the use of hitting in all of the 8 situations presented).

The estimated effects on caregiver knowledge and attitude to discipline after 1 year in the program indicated a gain of 7 points in knowledge and a reduction of about 5 points for being in favor of the use of harsh discipline with children, as compared to caregivers in comparison villages. However, the effects after 2 years of participation in the RPP were reduced to a 6 point gain in knowledge and 3 points decrease in favor of using corporal punishment. These losses were statistically significant (Table 3).

Changes in children’s language proficiency

Observed measures of children’s vocalizations. Children of caregivers in the RPP group showed impressive gains in how much language they produced during the play session (Figure 9). Although children in both groups increased the number of utterances they produced in 5 min at phase 2 compared to baseline – an unsurprising change since they had all aged by one year - the increase was 32% greater for children in the RPP group.
compared to those in the comparison group. The gain in number of child utterances in the 5 min play session remained unchanged after two years in the program – possibly matched to the amount of caregiver talk during the same session.

The estimated effect of the RPP on amount of child talk remained about the same after adjusting for key child factors, including age, gender, nutritional status, and baseline vocalizations. However, the estimated effect of the RPP program on number of child vocalizations and conversational turns per hour obtained from the all-day LENA™ audio-recordings was small and not statistically significant once adjusted for possible confounders (Table 4).

**Looking-While-Listening (LWL) procedure.** In Figure 10 we show the language processing trajectories from the LWL task for a subset of 59 older children who were 39-43 months at the phase 2 follow-up (or 27-31 months at baseline). We see that the children start at or near 50% accuracy and look at the target picture for an increasing proportion of time as they process and understand the target word, as we would expect. However, children in both program and comparison groups demonstrated nearly identical trajectories, with overlapping error bars (Figure 10).

![Figure 10](image_url)

*Figure 10.* The time course of looking to the target picture as a function of time in msec from the onset of the critical word for children in the RPP Program and Comparison groups, after one year.

In addition, we failed to find a significant difference between groups on the summary measures of mean accuracy and reaction time from the 1-year and 2-year regression effect estimates (Table 4). As a final check, and to gain more statistical power, we combined data from phase 3 RPP children with the same-age older cohort from phase 2 in RPP villages and compared their summary LWL measures with those of the older cohort from phase 2 in comparison villages. The differences remained statistically
insignificant. It is important to note, however, that we had a large amount of variation among children in their mean accuracy and reaction times (Table 2) and we were not able to explain most of this variation in the regression analyses, despite including a large number of potential explanatory factors. Although the overall trajectory of language processing was what we would expect from young children performing the LWL task, the data suffered from a substantial amount of unexplained variance (adjusted R-squared=0.19 for RT and 0.14 for accuracy).

Table 4: Adjusted\(^a\) RPP program effect estimates for key child outcome measures

<table>
<thead>
<tr>
<th></th>
<th>1-year Effect estimate, (\Psi^i) (\beta) (SD)</th>
<th>2-year Effect estimate, (\Psi^ii) (\beta) (SD)</th>
<th>Additional year estimate (\Psi^iii) (\beta) (SD)</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Child Outcomes: All Ages(^c)</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Play session: Child utterances</td>
<td>16.6 (2.9)***</td>
<td>16.2 (4.2)***</td>
<td>-0.4 (3.9)</td>
</tr>
<tr>
<td>Child language milestones</td>
<td>2.2 (0.4)***</td>
<td>3.7 (0.6)***</td>
<td>1.5 (0.8)**</td>
</tr>
<tr>
<td>All-day audio: Child vocalizations/hr</td>
<td>15.2 (19.2)</td>
<td>-18.8 (31.6)</td>
<td>-33.9 (22.4)</td>
</tr>
<tr>
<td>All-day audio: Conversational turns/hr</td>
<td>2.5 (5.2)</td>
<td>-4.1 (8.6)</td>
<td>-6.6 (6.7)</td>
</tr>
<tr>
<td><strong>Child Outcomes: 20-32 m at Baseline(^d)</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>CDI: expressive vocabulary</td>
<td>6.9 (1.9)***</td>
<td>12.5 (2.5)***</td>
<td>5.6 (3.1)*</td>
</tr>
<tr>
<td>LWL Accuracy</td>
<td>0.001 (0.019)</td>
<td>-0.007 (0.027)</td>
<td>-0.009 (0.022)</td>
</tr>
<tr>
<td>LWL Reaction Time</td>
<td>-16.7 (34.6)</td>
<td>24.2 (42.6)</td>
<td>41 (33.3)</td>
</tr>
</tbody>
</table>

\(^a\) Covariates are age-adjusted lagged dependent variables plus characteristics of the child (age, change in age, age squared, gender, length-for-age, birth size), caregiver (age, relation to child, education, literacy, attended earlier Tostan CEP, children < 5 years in her care, lost a child, depressive symptoms), household (polygynous head of household, members who attended earlier Tostan CEP, mean age, median years of education, children < 5 years, crowding, wealth), and village (population, households, midwives, and marabouts; health clinic, primary school, day care time to market and to get water, electricity, operational NGOs, flooding in 2013).

\(^b\) Difference in effect estimates between one and two years of exposure

\(^c\) Caregiver-child play session: Child utterances; Caregiver report questionnaire data: Child language milestones; All-day recording using LENA\(^TM\): Child vocalizations, conversational turns.

\(^d\) MacArthur-Bates CDI adapted “understands and says:” expressive vocabulary; Looking-while-listening task: accuracy, reaction time.

\(\star\) p-value <.1, ** p-value <.05, *** p-value<.01
Parents’ report of children’s language proficiency. We found significantly greater gains in language milestones and expressive vocabulary for children in the RPP villages after one year in the program than for children in the comparison villages (Table 4). The number of language milestones achieved and vocabulary size continued to grow as the children aged by another year. Although the gains that children in the RPP villages experienced as compared to controls were small, they compounded over time, such that the difference between the phase 3 RPP children and the same-age phase 2 children in the comparison villages was larger than the first year effect estimate.

Figure 11. Linear predictions of milestones achieved with age using a simple regression model. The red line represents the change in milestones as a function of age for all children at baseline (RPP and controls, dashed red) and the controls alone one year later (solid red). The blue line represents the change in milestones from the cross-section of different aged children in the RPP villages after one year in the program. The green line represents the change from the cross-section of different aged children in the RPP villages after two years in the program. The solid (first year) and dashed (second year) black lines represent the longitudinal growth of the average 4 month old child.
In Figure 11, we use linear predictions from a regression model to demonstrate the change in milestones achieved with increasing age and between program and comparison groups. The one-year program effect estimate was approximately 2 milestones for children of all ages. We tested for, but did not find, a significant difference in program effect by child age. Therefore the cross-sectional age trend at baseline (blue line) is parallel to the cross-sectional age trend at phase 2 follow-up (red line) indicating the same 2-point improvement for all ages. After two years in the program, we still did not find an age by RPP program interaction. The estimated RPP effect at two years is about 4 points for all ages of children: the cross-sectional age trend at phase 3 follow-up (green line) is parallel to the red and blue lines.

The black lines demonstrate how the RPP effect appeared to double after two years. The lines represent the longitudinal growth of the average 4 month old child (as opposed to the cross-sectional change across different aged children). The solid black line links the score of a 4 month old child at baseline to their score one year later, post RPP. The dashed black line is the growth of the cohort of same-aged children after two years, and links the scores of a 4 month old from one year to two years post RPP.

If we were to replicate the black lines at each age of children at baseline, this would indicate how the cross-sectional lines are generated. Importantly, if we compare scores of RPP children when they reach 28 months of age, we find that the child who was 4 months at baseline will achieve about 2 points more gain in milestones than the RPP child who was 16 months at baseline. These results reflect the conventional wisdom of the importance of starting to talk to children early. However, the benefit of the first year of the RPP may have had this same long-term effect, even in the absence of the second year of RPP activities.

**Section 6 - Discussion**

This evaluation study demonstrated that Tostan’s parent-education program in rural Senegal successfully encouraged caregivers to talk more with their young children, which in turn was associated with greater gains in their children’s language development. We found that caregivers who had participated in the program dramatically increased the amount of verbal engagement with their children in one-on-one play sessions when observed one year later, and that they maintained a high level of engagement after a second program year. Caregivers in the comparison group showed no change in the play sessions over the first year period. In the same play sessions, children of caregivers in the RPP program produced more utterances after one year as compared to their comparison group peers. Finally, children in the program villages showed greater gains in language milestones and expressive vocabulary after one year, with long-term compounding gains for younger children after two years.

The large increase in child-directed speech among caregivers in the RPP group indicates that Tostan was very successful in teaching caregivers new ways to verbally engage with their young children, and that the children responded positively with increased speech. However, all-day recordings from the LENA™ technology showed no significant improvement in the amount of adult verbal engagement with children in the more “real world” situation of multiple caregivers. In fact, we found that the mean number of
adult female words per hour spoken in the presence of the child dropped from baseline to the follow-up one year later, in both the program and comparison villages, and that this downward trend continued in the program villages during the second year. These findings are consistent with what others have reported: that children in African villages spend less time in the presence of adult caregivers as children age and become more mobile (LeVine et al., 1996).

In addition, the gains in caregiver-report measures of children’s language development were not reproduced with the direct measures of children’s language processing efficiency obtained from the looking-while-listening (LWL) task. Improving our power to detect an effect on the LWL measures by combining same-age children at both follow-up periods did not change this result. The absence of a program impact on language processing efficiency may be due to the comparable levels of children’s all-day, every day verbal engagement with adults across program groups. However, we cannot discount the possibility that the absence of impact may be an inability to detect an effect due to unexplained variance in our LWL measures from either a lack of familiarity with the task or non-ideal conditions in our testing sites.

One of the key goals of the second year of the program was for Tostan to gain the support of religious leaders and fathers in promoting change in caregiver behavior with infants and young children. In addition, Tostan implemented a sustainability component to the program, in which they trained a small group of caregivers in each village to be experts in the RPP skill set and to continue with the home visits during the second year. However, the results of the second program year indicated that no additional skills were gained by the caregivers in our sample, but rather we found either a ceiling effect (i.e., for amount of caregiver and child talk in the one-on-one play session) or a slight loss of effect (i.e., for knowledge and attitude toward discipline). The main improvement seen after the second year was the gain in caregiver-reported language abilities for children who started the program at an earlier age.

We conclude from these results that Tostan brought about significant change in the behavior of primary caregivers. But the caregiver-child interactions with the RPP participant who had learned about the importance of child-directed speech in the program may have been limited – either because children spent less time in their caregiver’s presence as they aged or the caregiver did not incorporate what she learned into her everyday life. Given the lack of improvement following the second year of broad community-level information dissemination, the challenge remains for Tostan to identify how to increase the amount of rich adult verbal engagement that young children experience in these settings. This likely involves providing RPP training to all of the child’s caregivers – including their older siblings and grandparents – as well as improving the training to include new ways of incorporating child-directed talk while doing chores and activities throughout an ordinary day.

Our study was subject to several limitations. While the Tostan RPP intervention was conducted in 200 villages across five regions of Senegal which differed in language and cultural traditions, our evaluation was limited to 24 Wolof villages in the Kaolack region. Thus we don’t know if our positive findings on the effect of the RPP on amount of verbal engagement with young children can be generalized to the other language and cultural groups in which the program was conducted. And because Tostan had pre-selected villages to receive the RPP program, random assignment was not possible. Thus we carefully
matched comparison villages with RPP villages and adjusted statistically for measured confounders. Although caregivers may have differed on unobserved characteristics, the comparability of the relevant measures at baseline across the two groups was reassuring.

An additional limitation was that although we went to great lengths to use direct measures of both caregivers’ and children’s behaviors, we also relied on standard parent-report measures that are inherently subject to over-reporting bias. To address this concern, interviewers frequently requested examples of when the child was reported to say a word or do an activity. We also found positive correlations between these parent-report measures and the direct measures of child vocalizations from the all-day recordings, providing evidence of validity. Note that even direct observations in the caregiver-child play session are also potentially subject to bias. For example, caregivers in the RPP group might have talked more to their child at follow-up if they were aware of Tostan’s goals or more at ease with the research team. But since the STEP team was not involved in the Tostan program and spent equal amounts of time as welcome guests in both RPP and comparison villages, it is unlikely that their observations a year later were influenced by these factors. Moreover, the substantial increase in caregiver talk from baseline to follow-up in the RPP villages was accompanied by a 32% increase in children’s language production in the play session, a large effect that cannot be explained by demand characteristics of the observation.

Despite these limitations, the results of the RPP evaluation represent an impressive achievement for Tostan. Senegalese mothers with no formal education, living in subsistence-level rural villages, were motivated to learn new ways of interacting with their young children, parenting practices which were often inconsistent with prevailing social norms and traditional beliefs. Tostan’s approach was never to challenge such traditional beliefs directly, but rather to provide participants with alternative explanations, by sharing scientific evidence about the timing of brain development and the ways in which children benefit from a cognitively stimulating environment. In this way, Tostan aimed to remove the social stigmas associated with talking to and engaging with babies. The aspirations that parents discussed in the focus groups for their children and their children’s future will be important for keeping parents involved and thus improving and sustaining Tostan’s parental education program going forward.

References
doi.org/10.1177/0741713610363017