NBER WORKING PAPER SERIES

LEVERAGING ARTIFICIAL INTELLIGENCE AND FIELD EXPERIMENTS TO EXPLORE NOVEL FEATURES OF PARENTAL SPEECH AND FOSTER CHILD DEVELOPMENT

Julie Pernaudet
John A. List
Arnoldo Müller-Molina
Majid Ahmadi
Imrul Huda
Ajay Sailopal
Dana Suskind

Working Paper 35302
http://www.nber.org/papers/w35302

NATIONAL BUREAU OF ECONOMIC RESEARCH
1050 Massachusetts Avenue
Cambridge, MA 02138
June 2026

The studies received approval from the University of Chicago Institutional Review Board . We are grateful to Amanda Bezik, Christy Leung, and Michelle Saenz for the implementation and monitoring of the projects. We also thank participants of the ASSA 2023 meeting, the ESA 2023 World Meeting, the Synapse Lab seminar at the University of Stavanger, the AFE 2023 Conference, the 2023 Kumar Conference on Early Childhood Development, the UChicago Committee on Education workshop, the 2024 workshop on Childhood Education at NYU Abu Dhabi, the 2024 ESIF Economics and AI+ML Meeting, the 2024 SREE Conference, and the 2024 AI in Social Science Conference for their useful comments. The two studies used in this paper were funded by the PNC Foundation and the Heising-Simons Foundation. The views expressed herein are those of the authors and do not necessarily reflect the views of the National Bureau of Economic Research.

NBER working papers are circulated for discussion and comment purposes. They have not been peer-reviewed or been subject to the review by the NBER Board of Directors that accompanies official NBER publications.

Leveraging Artificial Intelligence and Field Experiments to Explore Novel Features of Parental Speech and Foster Child Development

Julie Pernaudet, John A. List, Arnoldo Müller-Molina, Majid Ahmadi, Imrul Huda, Ajay Sailopal, and Dana Suskind

NBER Working Paper No. 35302
June 2026
JEL No. C45, C93, I24

ABSTRACT

Parents play a critical role in shaping children's skills during the first years of life. Yet, identifying the contributors to richer learning environments remains difficult due to various unobservable factors. In this paper, we combine field experiments with AI to explore new acoustic features of parental speech. Specifically, we develop a signal processing model that uses more than 600 hours of recorded parent-child interactions combined with assessment data from two home-visiting experiments conducted in the Chicagoland area to identify features of parental speech that map into children's skills. Our two experiments consist of the same intervention helping parents provide nurturing interactions to their child. We exploit the experimental and natural variation in our data to explore two causal channels and one potential moderator. First, our intervention improves parental speech consistently across the two studies, as measured by acoustic features that are predictive of higher socioemotional skills and adult-child conversational turns. Further, we find that it also increases children's language skills in both experiments, as well as socioemotional skills in the second experiment. Interestingly, our heterogeneity analyses reveal that some of the interventions' impacts vary by socioeconomic groups, with patterns across the two experiments suggesting that the mechanisms are context-dependent.

Julie Pernaudet
The University of Chicago
Department of Economics
julie.pernaudet@gmail.com

John A. List
The University of Chicago
Department of Economics
and Australian National University
and also NBER
jlist@uchicago.edu

Arnoldo Müller-Molina
arnoldomuller@gmail.com

Majid Ahmadi
The University of Chicago
majid.ahmadi@gatech.edu

Imrul Huda
imrul66@gmail.com

Ajay Sailopal
ajays@uchicago.edu

Dana Suskind
The University of Chicago
Biological Sciences Division (BSD)
dsuskind@surgery.bsd.uchicago.edu

A randomized controlled trials registry entry is available at Registered on clinical trials, see paper for urls

Introduction

Babies born today face four inextricable facts: i) their experiences in the first few years of life will be key determinants of how the next 75 turn out (1-5); ii) the family and community that they are born into will govern and moderate those early experiences, leading to considerable socioeconomic disparities (6-10), iii) knowledge of the causal pathways that determine these outcomes remains an opaque puzzle; and iv) recent technological and scientific advances provide a unique opportunity to shed light into this puzzle by detailing the inter-relationships between factors hypothesized to contribute to early human development.

These four facts represent our starting point to begin an exploration of early skill formation and its relation to parental inputs. A key to unlocking human potential and lessening socioeconomic disparities is a deep understanding of the production of early skills. To develop our framework, we build on recent advances in machine learning and leverage two home-visiting field experiments. The field experiments recognize that one key feature of a newborn's environment is auditory signals; indeed, the first salient auditory component babies are exposed to is the pitch of the caretaker's voice (11).

We are not in virgin territory with our choice to explore auditory signals.[1] The origin of the science of acoustics is generally attributed to Pythagoras, with modern studies typically having roots in the work of Galileo. As Titze et al. (2015) note (13), "the study of vocalization brings together a long history of voice terminology from acoustics, linguistics, phonetics, speech pathology, laryngology, music, theater, biology, and speech technology." Our study introduces economics and predictive artificial intelligence (AI) to this admirable group by creating a model that generates novel metrics for the analysis of parental speech. To do so, we create a multi-step algorithm that characterizes parental voice by its specific acoustic features, and we build speech indices that reflect the predictive power those features carry for different child outcomes.

To obtain exogenous variation in those acoustic features, we leverage two randomized field experiments that promote responsive parenting among low-income families through a six-month-long home-visiting program. The first experiment focuses on English-speaking families with children aged 13-16 months at enrollment and followed over four years, while the second experiment focuses on Spanish-speaking families with children aged 24-30 months at enrollment and followed over one year. Both studies were conducted in the Chicagoland area among low-income families. The randomization samples were respectively 206 and 91 families in Exp.1 and 2. As we followed families over time, we collected more than 600 hours of audio recordings of parent-child interactions and conducted regular assessments of children's cognitive and non-

[1] Researchers argue that the evolutionary origin of vocal communication dates back more than 400 million years (12)

cognitive skills. This rich dataset allows us to build bridges between several fundamental fields of research in early childhood; the fast-growing research on the use of neural networks for analyzing audio recordings of children's linguistic environments, research in developmental psychology and neuroscience on the importance of parental voice features (e.g., *parentese* or *baby talk*) for brain development, as well as recent research on the issue of replicability and scalability in early childhood development. We discuss those different areas of research in more detail at the end of the introduction and in the Materials and Methods section.

To put our field experimental data in motion, we extract a large set of speech features that we map into scores capturing the language and socioemotional skills of children. More specifically, we predict weights for all features based on their correlation with each type of skill and create speech indices that correspond to the weighted sum of the features. We conduct this estimation on the control group data to avoid capturing the potential effect of the intervention, which could confound the correlation, and extrapolate it to the full sample. Unlike more commonly used approaches that involve manually tagging audio datasets and training models to predict those tags, we take advantage of our child outcome measures to adopt a data-driven approach that allows the model to agnostically determine which speech features matter the most. While this approach does not yield metrics that are directly observable and interpretable, it can be easily reproduced among broader populations. Another key advantage of our field experiments is that the experimental variation allows us to assess the malleability of our speech indices. By analyzing parental speech and child outcomes in two separate field experiments, we are also able to assess the extent to which findings replicate or differ between the two populations.

Our combination of predictive AI and field experiments generates three areas of insight.

Our first key result is a causal insight gained from our two field experiments: the intervention shifts acoustic features of parental speech that, in our control sample, are predictive of higher child socioemotional outcomes, and increases the number of adult-child conversational turns per hour.. Second, we find that the intervention improves child outcomes, with increases in different measures of language skills across the two experiments and an increase in socioemotional skills in the second experiment. Last, our heterogeneity analysis reveals a complex and somewhat puzzling pattern that differs across our two studies. We find limited variability in impacts across education levels on parental inputs in both experiments, but striking differences in impacts on children's socioemotional skills, which increase substantially only among lower-educated families in the second experiment, while decreasing in the equivalent subgroup in the first experiment. Understanding why these patterns diverge represents an important direction for future research.

Overall, our combined approach contributes to the science of early education by offering novel measures of parental inputs that map into higher children's skills and are malleable. Because our

model is standardized and automated, it can be applied to other studies and used on a larger scale at no extra cost and while preserving the quality of the generated data (16).

We view our work as contributing to several branches of the literature. For example, it relates to the literature studying the role of parental inputs in the technology of early skill formation (see 17 for a theoretical framework and 18 or 19 for recent explorations of the role of parental beliefs about that technology). Numerous papers provide empirical evidence of the benefits of fostering parental investment for child development, but also longer-run outcomes (e.g., 3, 5, 20, 21). In particular, home-visiting programs targeted to low-income families have been shown to achieve large impacts on cognitive and socioemotional skills (18, 22-24). Our paper complements this research by generating new metrics of parental inputs that directly map into the technology of skill formation. By design, our speech indices give higher weights to features of parental speech that are most predictive of child outcomes. We show that our home-visiting program shifts those indices, thereby improving the speech inputs children receive from their parents.

Relatedly, there is extensive evidence, arising from developmental psychology and neuroscience, exploring how the quantity of parent-child interactions relates to children's outcomes (25-29). In addition, explorations of how it varies across socioeconomic backgrounds generally show that lower-educated parents provide fewer interactions (30,31). These studies typically rely on software such as the LENA technology (www.lena.org) to quantify interactions automatically. Assessing the characteristics of the voice of parents is more challenging, especially in large samples, as it usually requires human coding.

Adults tend to use specific voice patterns when interacting with young children, exaggerating pitch variation or vowel articulation, slowing down tempo, for instance, which is sometimes referred to as *parentese* or *baby talk*. Those speech patterns are known to facilitate children's learning (32-36) and exhibit a certain level of universality across cultures and languages (37-40). A series of articles by Naja Ferjan Ramírez, Patricia Kuhl, and co-authors, have shown that the use of parentese in early parent-child interactions predicts the language skills of children at later ages and parent coaching can help enhance that type of speech (41-44). Our contribution to this literature is to propose a novel, data-driven exploratory method to assess parental speech that does not rely on human coding of child-directed speech or other specific aspects of parental speech but instead uses child outcomes to agnostically determine which features of parental voice matter the most. While they do not translate into a specific speaking style, our metrics can be replicated and used in large-scale samples at a relatively low cost. As discussed in Lichand et al. (2022) (45), we believe the use of technologies integrating such models can help overcome the limitations of existing measurement tools and help address the replication crisis in early childhood research.

The remainder of the paper is organized as follows. The next section presents the impacts of the two experiments on parental speech and child outcomes, with an exploration of their heterogeneity by education levels, followed by a discussion of the main takeaways and limitations. In the Materials and Methods section, we present the experiments, the data we collected, and provide an overview of our AI approach. In the Supplementary Materials, we give a detailed description of the study design and sample, a comprehensive explanation of the different steps of our machine-learning model and construction of the indices, and we show the results of our robustness analyses.

Results

First, we investigate the malleability of our new measures of parental speech. More specifically, we test whether the home-visit program shifts our speech indices in a way that is conducive to skill formation.

Average impacts of the intervention on parental speech and child outcomes

In Figure 1, we show the two indices as well as the measure of conversational turns from LENA for each experiment. While the latter was part of the preregistered analysis plan, the former is a new exploration of the data.[2]

[Figure 1: ATEs of Interventions on Parental Speech]

The figure shows the results of a linear regression of each index on the binary variable indicating whether parents were randomly allocated to the treatment or control group. The coefficients provide the average treatment effects (ATEs) of the home-visiting program on the different dimensions of parental speech. For all metrics, the effects are standardized for easier comparison across outcomes within the study and with other studies. To assess the magnitude of the effects in their original unit, we refer the reader to Tables S.1. and S.2. in the Supplementary Materials, which provide the standard deviation of the different metrics used in the analysis for each experiment.

[2] The two experiments were preregistered on clinicaltrials.gov under the references NCT02216032 (https://clinicaltrials.gov/study/NCT02216032) and NCT03076268 (https://clinicaltrials.gov/study/NCT03076268).

Our estimation approach considers all time points in a panel model using random effects. That specification provides efficient variance estimates, but for the treatment effect estimates to be consistent, the treatment variable must be independent of the error term. While the randomization of parents across the two groups aims to fulfill that requirement, we observe some attrition over time that could induce observable and unobservable differences between the treatment and control groups. To address that concern, we also estimate a panel model with individual fixed effects that control for any time-invariant differences between the two groups. The results from the random effects model are presented in more detail in the Supplementary Materials, Section D, and the results from the fixed effects model are presented in Section E.1.

In both experiments, we find that the home-visiting intervention tends to improve our measures of parental speech. The index predicting language skills increases by 0.27 standard deviations in the first experiment (denoted Exp. 1 hereafter), and by 0.12σ in the second experiment (denoted Exp. 2 hereafter). Table S.9. in the Supplementary Materials indicates that the impact on the index predicting language skills has a p-value of 15% in Exp. 1 when accounting for the small sample size and multiple hypothesis testing (MHT), suggesting that this result should be interpreted with caution, and 4% in Exp. 2. If we now look at the second index, predicting socioemotional skills in children, we observe a positive increase of 0.11σ and 0.15σ in the first and second experiment, respectively. Both impacts are significant at the 1% level when adjusting for sample size and MHT.

Finally, the figure shows that the intervention further increases the number of conversational turns per hour by about 0.8 standard deviations, a result that replicates across experiments and was expected given that the home visitors set goals with the families in terms of conversational turns. Overall, our results suggest that our home visiting program consistently improves parental speech inputs in two different populations.

Examining Table S.13 in Section E of the Supplementary Materials, we can note that the positive impacts on the speech indices in the second experiment are slightly smaller and less precise when we control for time-invariant heterogeneity using a fixed effects model. In Exp. 1, the positive impact on the socioemotional index is unchanged, with an adjusted p-value of 2%. The impact on the language index is larger but still insignificant at standard levels when we adjust the p-value for sample size and MHT. Last, the impacts on conversational turns in both experiments are still large and significant at the 1% level.

Next, we turn to the analysis of child outcomes. Figure 2 compares again the average outcomes of the treatment and control groups, focusing on the language and socioemotional skills of the children.

[Figure 2: ATEs of Interventions on Child Outcomes]

In Exp. 1, language skills are measured by a z-score combining the PLS and PPVT assessments. In Exp. 2, those skills are measured by a z-score combining the ROWPVT and WM assessments. Socioemotional skills are measured in both cases using the ASQ-SE. All variables are standardized with respect to the mean and standard deviation of the control group distribution at the first assessment point where the measure was used. In cases where the first assessment point corresponds to the baseline, both groups are used, as the treatment group has yet to receive the intervention and is therefore comparable to the control group by virtue of randomization. Section A of the Supplementary Materials presents the list of measures used at each time point with an explanation for the acronyms used above.

In Exp. 1, Figure 2 reveals that the treatment effects are small and imprecise on both the language and socioemotional scores (blue bars). However, Table S.10 in the Supplementary Materials reveals that the absence of a significant impact on the language score masks important differences between the two measures used to assess children's language development. A first difference to note is the timing of those two assessments, with the PLS being implemented at each of the six time points and the PPVT being implemented only after children turned 2 years old, which corresponds to the fifth time point, a year and a half after the end of the intervention (see section A in the Supplementary Materials). This difference, due to differences in the age requirements of the two assessments, implies that we have substantially more statistical power to detect impacts with the PLS, including early impacts that might have faded by the time we conducted the PPVT assessments. Another difference between the two assessments is their content. The PLS provides a comprehensive play-based assessment of language skills, ranging from auditory comprehension to expressive communication skills, measured by gesture at early ages and language structure and semantics at later stages. The PPVT, on the other hand, is an assessment focused on receptive vocabulary, thereby capturing a narrower set of language skills. Table S.10 shows that while the impact on the PPVT is highly imprecise and not significant, the impact of the intervention on the PLS is large, with an average increase of 4.33σ. We should note that the magnitude of this impact partly reflects the increase in the PLS scores with age (see Table S.1. in the Supplementary Materials). As explained above, all scores in the analysis were standardized with respect to the first time point at which they were measured. Since the baseline PLS scores are low compared to the following time points, the magnitude of the standardized effects over the entire period is large.

Turning now to Exp. 2, we find significant improvements in both types of skills, the program leading to an increase of 0.30 standard deviations in language skills and 0.37 standard deviations in socioemotional skills. It should be noted that in Exp. 2, language skills are only assessed once, at

the last assessment point, hence the lower number of data points for that score and noisier estimate. For that specific outcome, we therefore use a simple OLS regression on cross-sectional data. The decomposition results presented in Table S.10 additionally show that the two measures used in the language score (WM and ROWPVT, respectively capturing expressive and receptive vocabulary) are similarly affected by the intervention, increasing by about 0.3σ.

Heterogeneous impacts of the intervention on parental speech and child outcomes

In this last section, we explore moderation, or whether the impact of the intervention on parental inputs and children's skills differs across socioeconomic subgroups. As participants in our studies all have at most a Bachelor's degree per our enrollment criteria, we make the distinction between parents whose highest level of education is a high school degree or below and parents whose highest level of education is above a high school degree at baseline, which is the split that maximizes the number of observations in the smallest subgroup across the two studies given the differences in the distribution of education levels in each sample. Those two subgroups represent about 30 and 70% of the sample, respectively, in Exp. 1 (see Table S.3 in the Supplementary Materials, Section B) and 70 and 30% of the sample, respectively, in Exp. 2 (Table S.4).

Using that binary indicator, we compare the treatment effects among parents whose highest level of education is a high school degree or below to the treatment effects among parents whose highest level of education is above a high school degree (referred to as conditional average treatment effects, CATEs). Figure 3 presents the same comparisons as Figure 1, split between the two experiments.

[Figure 3: CATEs of Interventions on Parental Speech (Exp. 1 in the top panel, Exp. 2 in the bottom panel)]

In Exp. 1, empirical results reveal that the positive average treatment effects found on the index predicting higher language skills in Figure 1 tend to be driven by lower-educated parents, although the difference in impacts between the two subpopulations is not statistically significant. The positive impact on the speech index associated with higher socioemotional skills is similar across the two subgroups and significant at the 5% level in both cases (see Table S.11 in the Supplementary Materials). By contrast, in Exp. 2, the positive effect on the socioemotional index is three times larger among lower-educated parents (0.18σ versus 0.06σ), but the lack of precision of the estimates does not allow to reject the hypothesis that this difference is null. Last, the figure shows

that the impact of the intervention on the CTCs tends to be higher on more educated parents in both studies, but here again, that difference is not statistically significant.

In section E.2. of the Supplementary Materials, we reproduce the heterogeneity analysis with equalized subgroup sizes by randomly subsampling the larger group. Although we find similar patterns in general, the estimations remain imprecise and difficult to interpret (see section E.2. for a more detailed discussion).

Overall, the heterogeneity analysis reveals some nuanced differences in impacts between socioeconomic subgroups, but the studies are not powered to draw robust conclusions. We see those results as only suggestive and requiring further exploration in larger studies.

Lastly, we compare the impacts of the intervention on child outcomes across the two socioeconomic subgroups. The results are presented in Figure 4.

[Figure 4: CATEs of Interventions on Child Outcomes (Exp. 1 in the top panel, Exp. 2 in the bottom panel)]

In Exp. 1, impacts on language skills are highly imprecise when combining the PLS and PPVT measures, but here again, this result masks large differences between the two measures, with the PLS showing increases of similar magnitudes in both subgroups (around 4.1-4.2σ). On the other hand, socioemotional skills seem to be negatively affected among children coming from lower-educated families.

Exp. 2 data show that the impact of the intervention on language skills is similar across both subgroups (around 0.3σ) but noisy. Interestingly, the heterogeneity analysis reveals that the positive impact of the home-visiting program on socioemotional development is entirely driven by children coming from lower-educated families, who experience an increase of 0.55σ in skills. The impact is significantly higher than the impact among higher-educated families, which is not statistically different from zero, and robust to multiple hypothesis testing adjustment, even when we restrict the analysis to the bootstrapped sample that equalizes the subgroup sizes (see Tables S.12 and S.17).

This last result has potentially important implications, as it shows that the home-visiting program reduces disparities in an important skill for early child development among the population targeted in the second experiment. Additional analyses of the pre-intervention data (Table S.18) indeed show that at age 24-30 months, children of lower-educated parents score 0.22-0.24σ lower than children coming from higher-educated backgrounds on our socioemotional measure. That socioeconomic difference flips signs right after the intervention in the treatment group, with children from lower-educated backgrounds scoring higher on the ASQ-SE than children from higher-

educated backgrounds, while the initial socioeconomic gap doubles in the control group. Six months post-intervention, the gap is not significantly different from zero in the treatment group and has almost tripled to reach 0.64σ in the control group. Our intervention thereby seems to prevent early inequalities in socioemotional skills from magnifying as children grow in the second experiment.

Discussion

Research on child development across the spectrum of the sciences is unequivocal: parental inputs play a fundamental role in shaping children's learning trajectory. Yet, we are equipped with surprisingly limited tools to measure and analyze such inputs, confining empirical work on skill formation to small-scale observational data that require extensive manual coding or larger-scale survey data that provide only coarse measures of parental investments. This paper attempts to fill this gap by leveraging the opportunities offered by artificial intelligence to process large amounts of data at a granular level. We develop an acoustic model that allows us to analyze more than 600 hours of audio recordings of parent-child interactions and explore new channels for how parents' talk affects early skill development.

Using data from two field experiments as input to the AI model, we conduct an exploratory analysis of the ways acoustic patterns of speech may be altered by our intervention. Our first findings show that those speech features are positively impacted by the intervention. A simple home-visiting program promoting richer interactions based on the 3Ts framework (Tuning in, Talking more, and Taking turns) leads to changes in speech features that are predictive of higher child outcomes consistently across the two studies. Additionally, we find that the program leads parents to engage in more interactions with their children in both studies, as measured by an increase in the number of conversational turns per hour. While the latter was an explicit objective of the intervention, the impacts we observe on speech features show that the program can also positively affect other (less directly observable) aspects of parental speech, and thereby improve parent-child interactions more broadly.

A decomposition of those average treatment effects by socioeconomic subgroups suggests potential differences in the impact of the intervention on the speech indices and conversational turns, but those differences are imprecisely estimated and sensitive to the model specification. Going one step further, we find that socioeconomic gaps in children's socioemotional skills shrink with the intervention in the second experiment, while the first experiment, by contrast, reveals a negative impact on socioemotional skills among low-educated families. Several factors could contribute to these divergent patterns. First, the two experiments differ in child age at enrollment (13-16 months vs 24-30 months), and developmental timing may moderate how the same program

affects skill development. Second, the linguistic and cultural contexts differ substantially between English- and Spanish-speaking populations in our samples, which could affect both baseline patterns and intervention mechanisms. These inconsistencies prevent us from drawing firm conclusions about whether and how the intervention reduces socioeconomic disparities. Overall, our heterogeneity analysis reveals that the intervention's impacts tend to vary across socioeconomic groups, but the mechanisms are context-dependent in ways we cannot fully characterize with the current data and require further research.

Although we view those findings as a first step towards the discovery of novel ways in which parental inputs affect child development, we also want to reiterate that our analyses are exploratory and encourage future work to try and replicate them in different studies. Various factors enter into the production of early skills; parental speech patterns are one of them, but further investigations should be conducted to better understand how they relate to other important determinants of child development that we do not observe in our data. We should also note that our estimation of the mapping between parental speech patterns and child outcomes is correlational. Our analysis shows that the home-visiting intervention causally moves conversational turns, parental speech patterns, and children's skills, but it was not designed to assess the causal relationship between those three types of outcomes. Another limitation of our analysis is the multiple differences between our two studies (e.g., target population, children's age, assessments, timeline), making it difficult to identify what exactly may cause the similarities and discrepancies we observe in the findings. Further research needs to be conducted to identify the specific role played by sociocultural differences between the two populations.

In closing, and relatedly, we would be remiss not to mention the generalizability of our results and additional research directions. In terms of the former, we follow the SANS conditions in List (2020) (46) to understand the generalizability of our results. Our sample is a selected subset of individuals who agreed to participate in our experiment. In this sense, whether the results generalize to other populations is an open question, yet we view our primary contribution as providing a tool and framework to capture what was previously thought to be unmeasurable. In terms of attrition, once a subject started, very few opted out. Considering the naturalness of the choice task, setting, and time frame, we use a framed field experiment; thus, our setting is one in which individuals are engaged in a natural task and margin.

Finally, since we view our demand side results as WAVE1 insights, in the nomenclature of List (2020) (46), replications need to be completed to understand if our results can be applied to other populations. This point naturally leads to future research directions. Beyond replications, we urge researchers to more fully explore our tool to uncover other potentially important speech features that can be linked to child skill formation. In this spirit, we trust that scholars will engage in creating other mediation paths and causal moderators rather than stop at merely descriptive variants as we do in this study. Such explorations will help to detail what works and why for policymakers

interested in lessening childhood disparities. In addition, we trust that future work will tackle the interplay of the demand and supply sides of this market. Until key features are understood about the production and consumption sides, and relevant interactions, efficient policies to combat childhood inequities will be difficult, if not impossible.

Materials and Methods

Field Experiments

Intervention

The two field experiments used in the analysis are based on the same intervention, but they differ in terms of their target population and duration, as explained in the introduction. We provide more detailed information on the timeline of each study, the enrollment criteria, and the sociodemographic characteristics of each sample in the Supplementary Materials, Sections A and B.

The intervention consists of a series of twelve bi-monthly home visits conducted during the first six months after enrollment in both studies and aimed at helping parents create nurturing interactions with their children to foster their cognitive and non-cognitive development. The environment was a natural one, designed to be cost-effective and include any nuances faced at scale (47). The curriculum is centered around the "3Ts" framework – Tune in, Talk more, Take turns – with each of the twelve modules teaching parents different ways to use the 3Ts strategy in their everyday routines with their child (e.g., during book sharing, while preparing meals or playing). After each visit, the parent and home visitors set goals in terms of conversational turn counts (CTCs) that are assessed via LENA recordings of their daily interactions with their child. The detailed curriculum describing the content of each of the twelve modules can be found in Leung et al. (2020) (48, see Table 1, in particular). The curriculum was the same across the two studies, but the home visitors and implementation teams differed. For both experiments, families were randomized into a treatment group receiving the 3Ts home visits or into a control group receiving home visits centered on the nutritional needs of young children.

Data

The analysis relies on four types of data that we collected in both studies. The first data source arises from surveys completed by parents. In each case, parents provide information on the family's demographic characteristics and the socioemotional health of the child. We also collected data on parental inputs by filming parents while playing with and reading to their child during sessions of approximately 30 minutes (referred to as Free Play and Book Share in section A of the Supplementary Materials). We provided the same set of toys and books to all families to standardize the activities. Overall, we conducted seven video sessions in Exp. 1 and four video sessions in Exp. 2, representing more than 600 hours of recording in total. We use the video data for the speech analysis that we briefly describe in section 3 and explain in more detail in the Supplementary Materials, Section C. Further, we collected regular measures of the number of conversational turns per hour between the parent and child at home using the LENA technology. 2 The last type of data we collected consists of child assessment information for which specifically trained assessors who were blind to the treatment assignment conducted standardized tests with children to evaluate their language skills. In the first experiment, all assessments were conducted in English; in the second experiment, language assessments were conducted in Spanish and English and combined for the analysis. Assessments were conducted every six to twelve months. The two types of child outcomes our analysis focuses on are language skills (measured by the PLS and PPVT in Exp. 1, and the ROWPVT and Woodcock-Muñoz's Analogies and Picture Vocabulary subtests in Exp. 2) and socioemotional skills (measured by the ASQ-SE in both experiments).
We provide a complete description of the data collected at each time point for both studies in the Supplementary Materials, Section A.

AI approach

The model we develop aims to generate new measures of parental speech inputs that map into children's skill formation. To accomplish this, we proceed in three main steps. First, we identify the segments of parental speech in our audio recordings and isolate them from children's speech or other nonhuman sounds. This step is what we refer to as the speaker labeling model. For this, we use Pretrained Audio Neural Networks (PANN) that we adapt to predict our three classes of interest - mother, father, and child. Our model reaches precision levels between 0.64 and 0.72 for our main speaker of interest (mothers, who represent 97% of the parents in Exp. 1 and 100% in Exp. 2) when compared to validated human transcriptions. Next, we characterize parental speech by extracting a wide range of frequency features such as chroma features, contrast features, Mel-Frequency Cepstral Coefficients (MFCCs), log-scaled Mel-spectrograms, spectral bandwidth, centroid, flatness features, and fundamental

frequency (F0) that we match with speaker labels at the second level. Features are extracted from the audio recordings using Python's *librosa* package. The last step consists of building indices that are linear combinations of the parental speech features and give higher weights to features that best predict children's language and socioemotional skills. In our two studies, MFCCs and middle to high-range Mel-spectrograms are often selected among the top predictors of both skills, with contrast and chroma features, reflecting the clarity and chromatic qualities of parents' speech, respectively, being important drivers as well. We elaborate on each of those steps in the Supplementary Materials, Section C.

In this manner, we view our work as speaking to the strand of research that uses speech processing techniques to identify specific acoustic features of child-directed speech. Piazza et al. (2017) (49), for instance, use MFCC features to build a classification algorithm showing that mothers alter their vocal timbre (or 'tone color') when communicating with their infants compared to other adults. Focusing on the fundamental frequency feature (f0), De Palma and VanDam (2017) (50) also find that mothers, and to a smaller extent, fathers, raise f0 in child-directed speech, thereby conveying higher-pitch sounds. Our approach relies on similar audio classification models to analyze the voice of our participants, but uses a novel approach that builds parental speech metrics based on their mapping with longer-term child outcomes. Databases used in the aforementioned literature usually contain audio recordings of parent-child interactions but not longitudinal data on children's skills. We also explore how parental speech relates to different types of children's skills. While most studies primarily focus on language skills, we additionally analyze the mapping with children's socioemotional health. Another strength of our data is the experimental variation generated by the randomization of participants to the home-visiting program, which allows us to test the malleability of our parental speech indices. To the best of our knowledge, the closest paper assessing the impacts of a randomized parenting intervention on automatically coded acoustic features of parent-child interactions is Balsa et al. (2023) (51). Analyzing the fundamental frequency (f0) of adult vocalizations, they find that an e-messaging program promoting positive parenting practices results in parents using a larger pitch range. Our analysis contributes to those findings by including a broad range of potential features, as described in the Supplementary Materials, Section C.2, and linking them to child outcomes.

Acknowledgments

The studies received approval from the University of Chicago Institutional Review Board. We are grateful to Amanda Bezik, Christy Leung, and Michelle Saenz for the implementation and monitoring of the projects. We also thank participants of the ASSA 2023 meeting, the ESA 2023 World Meeting, the Synapse Lab seminar at the University of Stavanger, the AFE 2023 Conference, the 2023 Kumar Conference on Early Childhood Development, the UChicago Committee on Education workshop, the 2024 workshop on Childhood Education at NYU Abu Dhabi, the 2024 ESIF Economics and AI+ML Meeting, the 2024 SREE Conference, and the 2024 AI in Social Science Conference for their useful comments. The two studies used in this paper were funded by the PNC Foundation and the Heising-Simons Foundation.

References

1. Attanasio, O., Cortes, D., Maldonado, D., Rodriguez-Lesmes, P., Charpak, N., Tessier, R., ... & Gallegos, A. (2024). Parental Investments and Skills Formation During Infancy and Youth: Long Term Evidence From an Early Childhood Intervention (No. w32851). National Bureau of Economic Research.

2. García, J. L., Heckman, J. J., & Ronda, V. (2023). The lasting effects of early-childhood education on promoting the skills and social mobility of disadvantaged African Americans and their children. Journal of Political Economy, 131(6), 1477-1506.

3. García, J. L. and J. J. Heckman (2023). Parenting promotes social mobility within and across generations. Annual Review of Economics 15, 349–388.

4. Bailey, M. J., Sun, S., & Timpe, B. (2021). Prep School for poor kids: The long-run impacts of Head Start on Human capital and economic self-sufficiency. American Economic Review, 111(12), 3963-4001.

5. Gertler, P., J. Heckman, R. Pinto, A. Zanolini, C. Vermeersch, S. Walker, S. M. Chang, and S. Grantham-McGregor (2014). Labor market returns to an early childhood stimulation intervention in Jamaica. Science 344 (6187), 998–1001.

6. Chyn, E., & Katz, L. F. (2021). Neighborhoods matter: Assessing the evidence for place effects. Journal of Economic Perspectives, 35(4), 197-222.

7. Laliberté, J. W. (2021). Long-term contextual effects in education: Schools and neighborhoods. *American Economic Journal: Economic Policy*, *13*(2), 336-377.

8. Chetty, R., & Hendren, N. (2018). The impacts of neighborhoods on intergenerational mobility II: County-level estimates. The Quarterly Journal of Economics, 133(3), 1163-1228.

9. Pace, A., Luo, R., Hirsh-Pasek, K., & Golinkoff, R. M. (2017). Identifying pathways between socioeconomic status and language development. Annual review of linguistics, 3(2017), 285-308.

10. Kelly, Y., Sacker, A., Del Bono, E., Francesconi, M., & Marmot, M. (2011). What role for the home learning environment and parenting in reducing the socioeconomic gradient in child development? Findings from the Millennium Cohort Study. Archives of disease in childhood, 96(9), 832-837.

11. Liu, L., Götz, A., Lorette, P., & Tyler, M. D. (2022). How tone, intonation and emotion shape the development of infants' fundamental frequency perception. *Frontiers in Psychology*, 13, 906848

12. Jorgewich-Cohen, G., Townsend, S. W., Padovese, L. R., Klein, N., Praschag, P., Ferrara, C. R., ... & Sánchez-Villagra, M. R. (2022). Common evolutionary origin of acoustic communication in choanate vertebrates. Nature Communications, 13(1), 6089.

13. Titze, I. R., R. J. Baken, K. W. Bozeman, S. Granqvist, N. Henrich, C. T. Herbst, D. M. Howard, E. J. Hunter, D. Kaelin, R. D. Kent, et al. (2015). Toward a consensus on symbolic notation of harmonics, resonances, and formants in vocalization. *The Journal of the Acoustical Society of America 137* (5), 3005–3007.

14. Kalil, A., & Ryan, R. (2020). Parenting practices and socioeconomic gaps in childhood outcomes. The future of children, 30(1), 29-54.

15. Rowe, M. L. (2018). Understanding socioeconomic differences in parents' speech to children. Child Development Perspectives, 12(2), 122-127.

16. Al-Ubaydli, O., J. A. List, and D. Suskind (2020). 2017 Klein lecture: The science of us-ing science: Toward an understanding of the threats to scalability. *International Economic Review 61* (4), 1387–1409.

17. Cunha, F. and J. Heckman (2007). The technology of skill formation. American Economic Review 97 (2), 31–47.

18. List, J. A., J. Pernaudet, and D. L. Suskind (2021). Shifting parental beliefs about child development to foster parental investments and improve school readiness outcomes. Nature Communications 12 (1), 5765.

19. Cunha, F., I. Elo, and J. Culhane (2022). Maternal subjective expectations about the technology of skill formation predict investments in children one year later. Journal of Econometrics 231 (1), 3–32.

20. Carneiro, P., E. Galasso, I. L. Garcia, P. Bedregal, and M. Cordero (2024). Impacts of a large-scale parenting program: Experimental evidence from Chile. Journal of Political Econ-omy 132 (4), 000–000.

21. Attanasio, O., C. Meghir, and E. Nix (2020). Human capital development and parental investment in India. *The Review of Economic Studies 87* (6), 2511–2541.

22. Zhou, J., J. J. Heckman, B. Liu, and L. Mai (2024). The impact of a prototypical home visiting program on child skills. Journal of Labor Economics (forthcoming).

23. Attanasio, O., S. Cattan, E. Fitzsimons, C. Meghir, and M. Rubio-Codina (2020). Estimating the production function for human capital: Results from a randomized controlled trial in Colombia. American Economic Review 110 (1), 48–85.

24. Doyle, O. (2020). The first 2,000 days and child skills. Journal of Political Economy 128 (6), 2067–2122.

25. Huber, E., N. M. Corrigan, V. L. Yarnykh, N. F. Ramírez, and P. K. Kuhl (2023). Language experience during infancy predicts white matter myelination at age 2 years. Journal of Neuroscience 43 (9), 1590–1599.

26. Donnelly, S. and E. Kidd (2021). The longitudinal relationship between conversational turn-taking and vocabulary growth in early language development. Child Development 92 (2), 609–625.

27. Romeo, R. R., Leonard, J. A., Grotzinger, H. M., Robinson, S. T., Takada, M. E., Mackey, A. P., ... & Gabrieli, J. D. (2021). Neuroplasticity associated with changes in conversational turn-taking following a family-based intervention. Developmental cognitive neuroscience, 49, 100967.

28. Wang, Y., Williams, R., Dilley, L., & Houston, D. M. (2020). A meta-analysis of the predictability of LENA™ automated measures for child language development. Developmental Review, 57, 100921.

29. Zimmerman, F. J., J. Gilkerson, J. A. Richards, D. A. Christakis, D. Xu, S. Gray, and U. Ya-panel (2009). Teaching by listening: The importance of adult-child conversations to language development. Pediatrics 124 (1), 342–349.

30. Fibla, L., S. H. Forbes, J. McCarthy, K. Mee, V. Magnotta, S. Deoni, D. Cameron, and J. P. Spencer (2023). Language exposure and brain myelination in early development. Journal of Neuroscience 43 (23), 4279–4290.

31. Hart, E. R., S. V. Troller-Renfree, J. F. Sperber, and K. G. Noble (2023). Relations among socioeconomic status, perceived stress, and the home language environment. Journal of Child Language, 1–18.

32. Nencheva, M. L., E. A. Piazza, and C. Lew-Williams (2021). The moment-to-moment pitch dynamics of child-directed speech shape toddlers' attention and learning. Developmental Science 24 (1), e12997.

33. Adriaans, F. and D. Swingley (2017). Prosodic exaggeration within infant-directed speech: Consequences for vowel learnability. The Journal of the Acoustical Society of America 141 (5), 3070–3078.

34. Thiessen, E. D., E. A. Hill, and J. R. Saffran (2005). Infant-directed speech facilitates word segmentation. Infancy 7 (1), 53–71.

3541. Trainor, L. J. and R. N. Desjardins (2002). Pitch characteristics of infant-directed speech affect infants' ability to discriminate vowels. Psychonomic bulletin & review 9 (2), 335–340.

36. Werker, J. F. and P. J. McLeod (1989). Infant preference for both male and female infant-directed talk: a developmental study of attentional and affective responsiveness. Canadian Journal of Psychology/Revue canadienne de psychologie 43 (2), 230.

37. Hilton, C. B., C. J. Moser, M. Bertolo, H. Lee-Rubin, D. Amir, C. M. Bainbridge, J. Simson, D. Knox, L. Glowacki, E. Alemu, et al. (2022). Acoustic regularities in infant-directed speech and song across cultures. Nature Human Behaviour 6 (11), 1545–1556.

38. Parlato-Oliveira, E., M. Chetouani, J.-M. Cadic, S. Viaux, Z. Ghattassi, J. Xavier, L. Ouss, R. Feldman, F. Muratori, D. Cohen, et al. (2020). The emotional component of infant-directed speech: a cross-cultural study using machine learning. Neuropsychiatrie de l'Enfance et de l'Adolescence 68 (2), 106–113.

39. Kitamura, C., C. Thanavishuth, D. Burnham, and S. Luksaneeyanawin (2001). Universality and specificity in infant-directed speech: Pitch modifications as a function of infant age and sex in a tonal and non-tonal language. Infant behavior and development 24 (4), 372–392.

40. Grieser, D. L. and P. K. Kuhl (1988). Maternal speech to infants in a tonal language: Support for universal prosodic features in motherese. Developmental Psychology 24 (1), 14.

41. Ramírez, N. F., Y. Weiss, K. K. Sheth, et al. (2023). Parentese in infancy predicts 5-year language complexity and conversational turns. Journal of Child Language, 1–26.

42. Ferjan Ramírez, N., S. R. Lytle, and P. K. Kuhl (2020). Parent coaching increases conversational turns and advances infant language development. Proceedings of the National Academy of Sciences 117 (7), 3484–3491

43. Ferjan Ramírez, N., S. R. Lytle, M. Fish, and P. K. Kuhl (2019). Parent coaching at 6 and 10 months improves language outcomes at 14 months: A randomized controlled trial. Developmental Science 22 (3), e12762.

44. Ramírez-Esparza, N., A. García-Sierra, and P. K. Kuhl (2017). The impact of early social interactions on later language development in Spanish–English bilingual infants. Child Development 88 (4), 1216–1234.

45. Lichand, G., O. Leal Neto, J. Phuka, R. Chipojola, B. Laher, M. B. Enlow, A. E. Sidamon-Eristoff, K. Quigley, A. Weisleder, C. Lew-Williams, et al. (2022). The early childhood development replication crisis, and how wearable technologies could help overcome it.

46. List, J. A. (2020). Non est disputandum de generalizability? a glimpse into the external validity trial. Technical report, National Bureau of Economic Research.

47. List, J. A. (2024). Optimally generate policy-based evidence before scaling. Nature 626 (10102), 491–499.

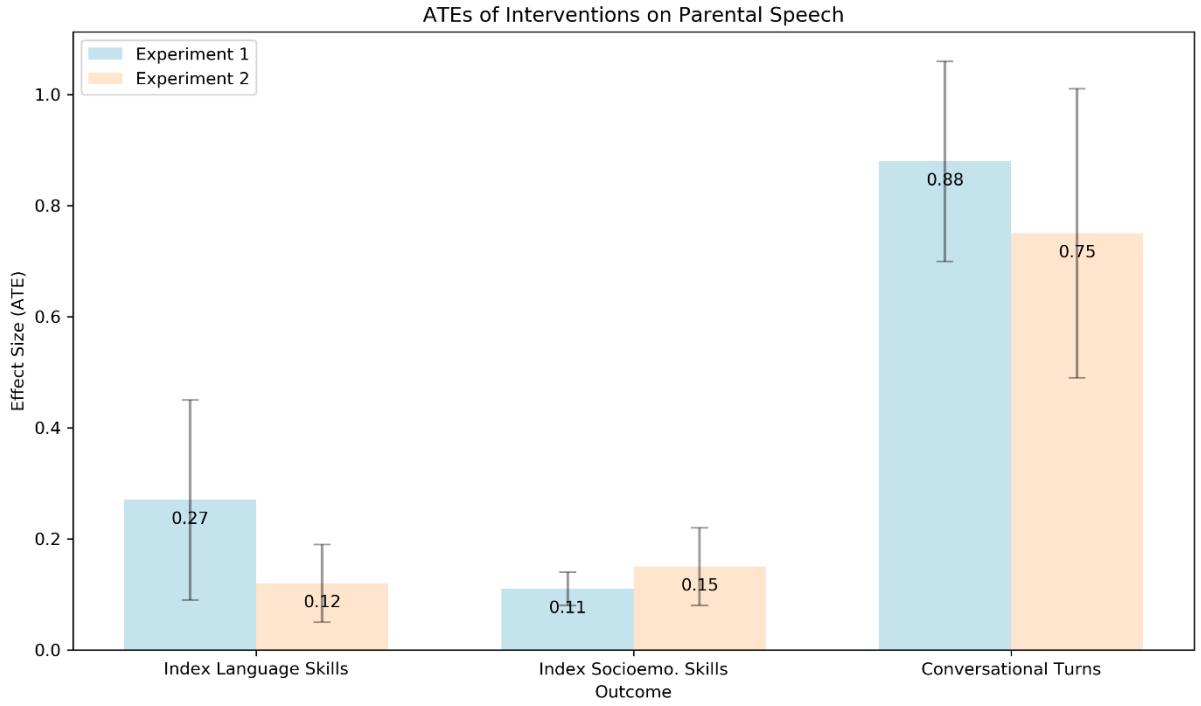
48. Leung, C. Y., M. W. Hernandez, and D. L. Suskind (2020). Enriching home language environment among families from low-SES backgrounds: A randomized controlled trial of a home visiting curriculum. Early Childhood Research Quarterly 50, 24–35.

49. Piazza, E. A., M. C. Iordan, and C. Lew-Williams (2017). Mothers consistently alter their unique vocal fingerprints when communicating with infants. Current Biology 27 (20), 3162–3167.

50. De Palma, P. and M. VanDam (2017). Using automatic speech processing to analyze fundamental frequency of child-directed speech stored in a very large audio corpus. In 2017 Joint 17th World Congress of International Fuzzy Systems Association and 9th International Conference on Soft Computing and Intelligent Systems (IFSA-SCIS), pp. 1–6. IEEE.

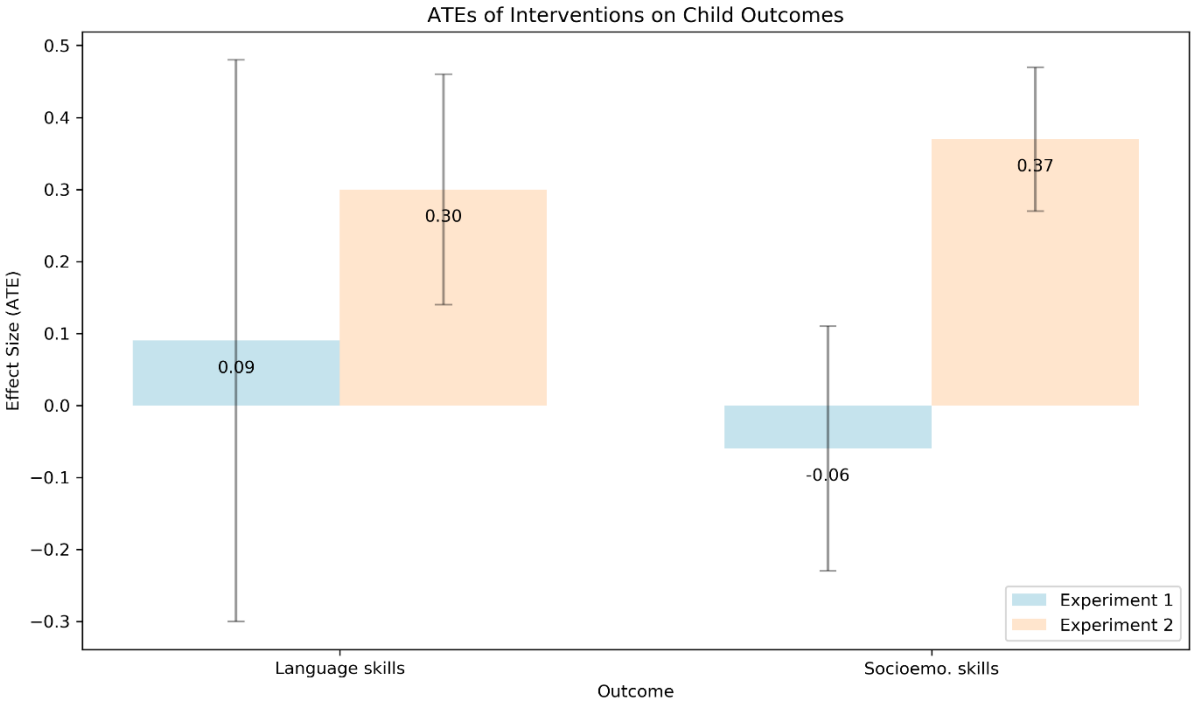
51. Balsa, A., F. L. Boo, J. Bloomfield, A. Cristia, A. Cid, M. de La Paz Ferro, R. Valdés, and M. del Luján González (2023). Effect of Crianza Positiva e-messaging program on adult–child language interactions. Behavioural Public Policy 7 (3), 607–643.

Figures



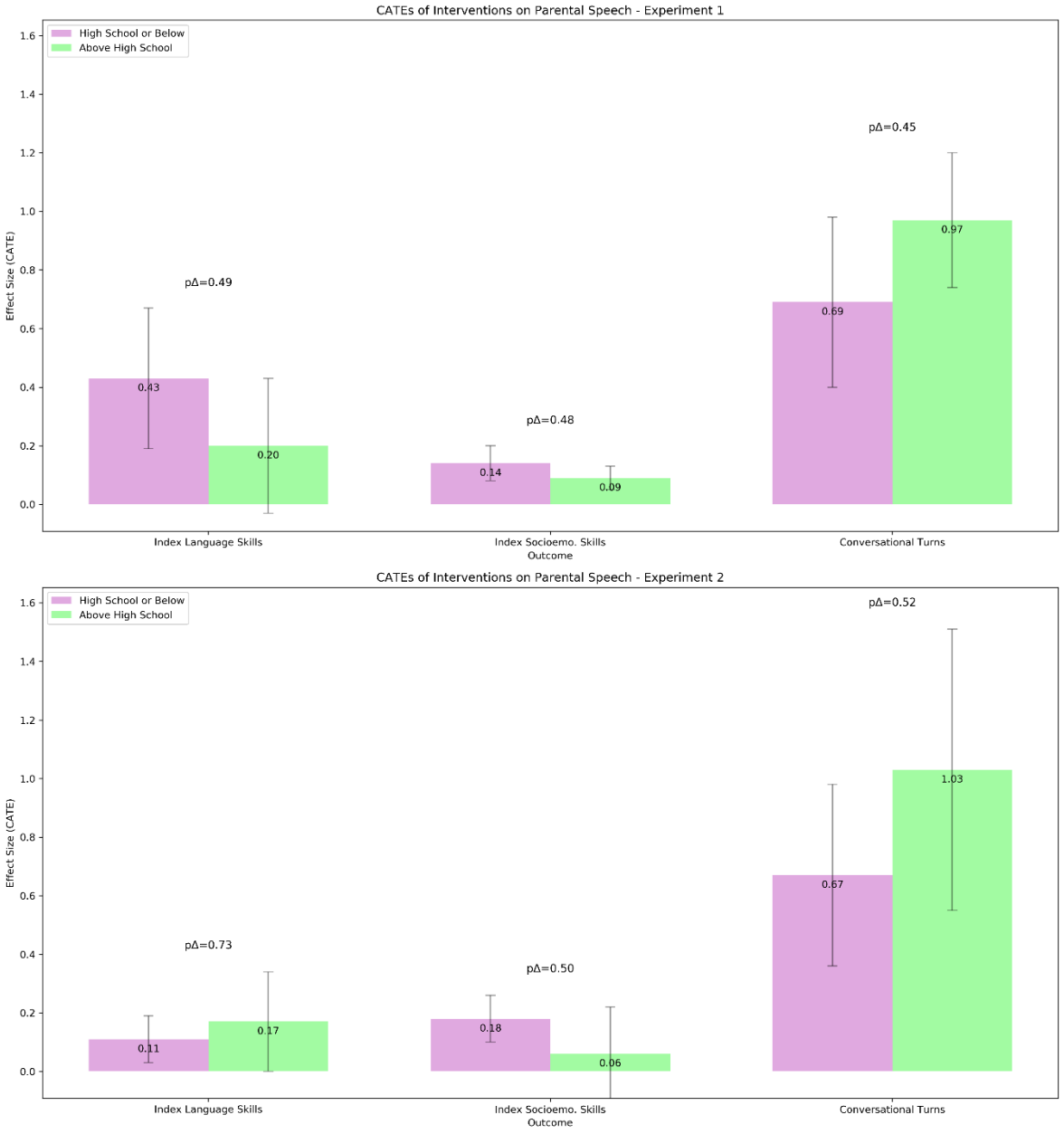
Note: Standardized differences between Treatment and Control groups with standard errors clustered at the individual level. Error bars show +/- the standard errors. Two-sided t-tests. Random effects model. Sample size: Experiment 1 - 861 for Conversational Turns, 962 for all others. Experiment 2 - 192 for Conversational Turns, 295 for all others. Sample size gives the number of observations in each panel regression, which is the number of families at each time point multiplied by the number of time points (see Figures S.1 and S.2 in the Supplementary Materials).

Figure 1: ATEs of Interventions on Parental Speech



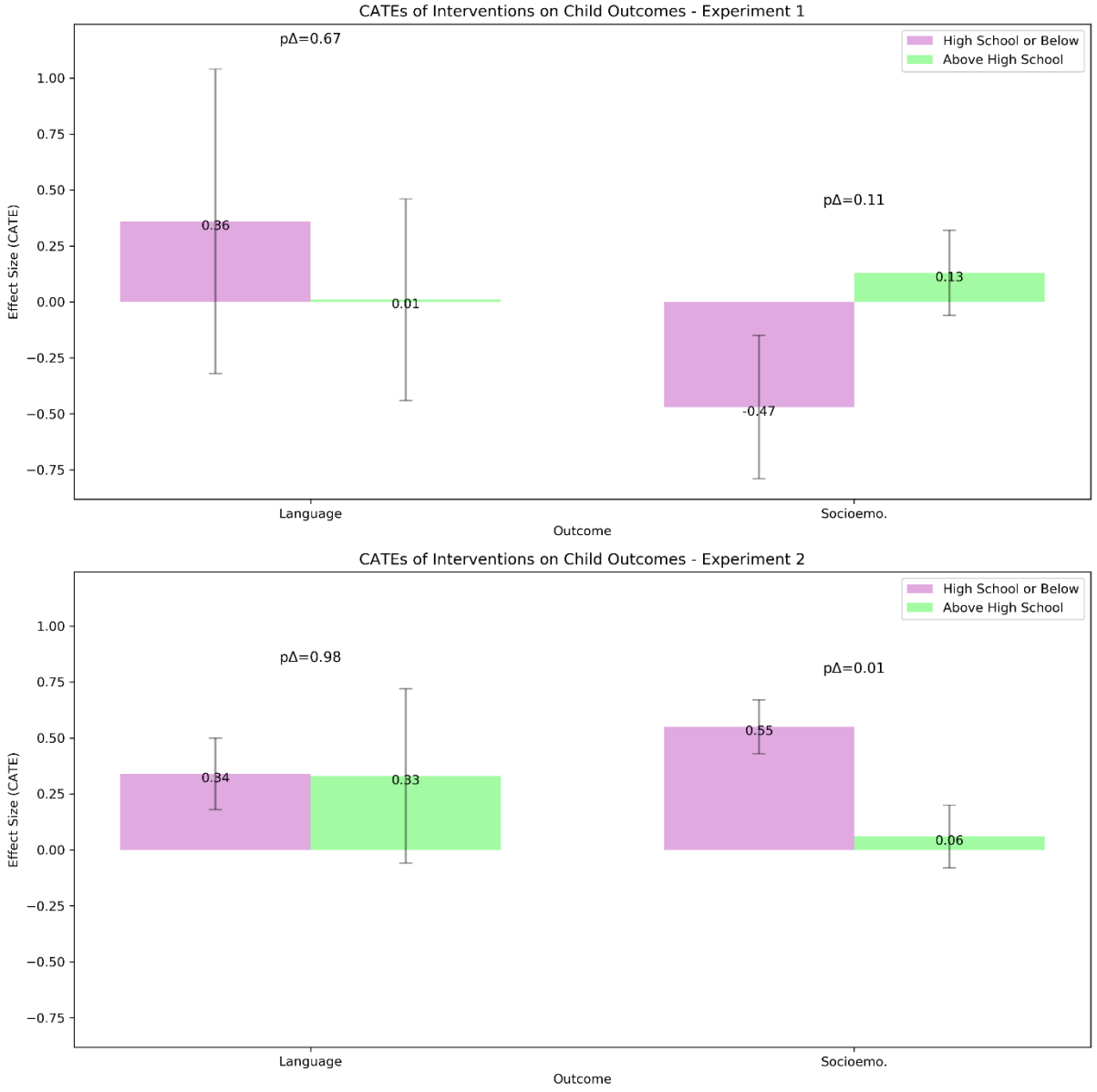
Note: Standardized differences between Treatment and Control groups. Random effects model with standard errors clustered at the individual level for panel data and simple OLS regressions with robust standard errors for cross-sectional data. Error bars show +/- the standard errors. Two-sided t-tests. Sample size: Experiment 1 - 258 for Language skills, 453 for Socioemo skills. Experiment 2 - 59 for Language skills, 216 for Socioemo. skills. Sample size gives the number of observations in each panel regression, which is the number of families at each time point multiplied by the number of time points (see Figures S.1 and S.2 in the Supplementary Materials).

Figure 2: ATEs of Interventions on Child Outcomes



Note: Standardized differences between Treatment and Control groups with standard errors clustered at the individual level. Error bars show +/- the standard errors. Random effects model. Pink bars give the treatment effect for parents whose highest level of education is a high-school degree or below. Green bars give the treatment effect for parents whose highest level of education is above a high-school degree. Two-sided t-tests. Above each pair of bars is the p-value of the test that the difference in CATEs between the two subgroups is null.Sample size: Experiment 1 - 861 for Conversational Turns, 962 for all others. Experiment 2 - 192 for Conversational Turns, 295 for all others. Sample size gives the number of observations in each panel regression, which is the number of families at each time point multiplied by the number of time points (see Figures S.1 and S.2 in the Supplementary Materials).

Figure 3: CATEs of Interventions on Parental Speech (Exp. 1 in the top panel, Exp. 2 in the bottom panel)



Note: Standardized differences between Treatment and Control groups. Random effects model with standard errors clustered at the individual level for panel data and simple OLS regressions with robust standard errors for cross-sectional data. Error bars show +/- the standard errors.Two-sided t-tests. Pink bars give the treatment effect for parents whose highest level of education is a high-school degree or below, Green bars give the treatment effect for parents whose highest level of education is above a high-school degree. Above each pair of bars is the p-value of the test that the difference in CATEs between the two subgroups is null.Sample size: Experiment 1- 258 for Language, 453 for Socioemo. Experiment 2- 59 for Language, 216 for Socioemo. Sample size gives the number of observations in each panel regression, which is the number of families at each time point multiplied by the number of time points (see Figures S.1 and S.2 in the Supplementary Materials).

Figure 4: CATEs of Interventions on Child Outcomes (Exp. 1 in the top panel, Exp. 2 in the bottom panel)

Supporting Information

The file includes:

A Study timeline and measures

Figure S.1: Timeline Experiment 1

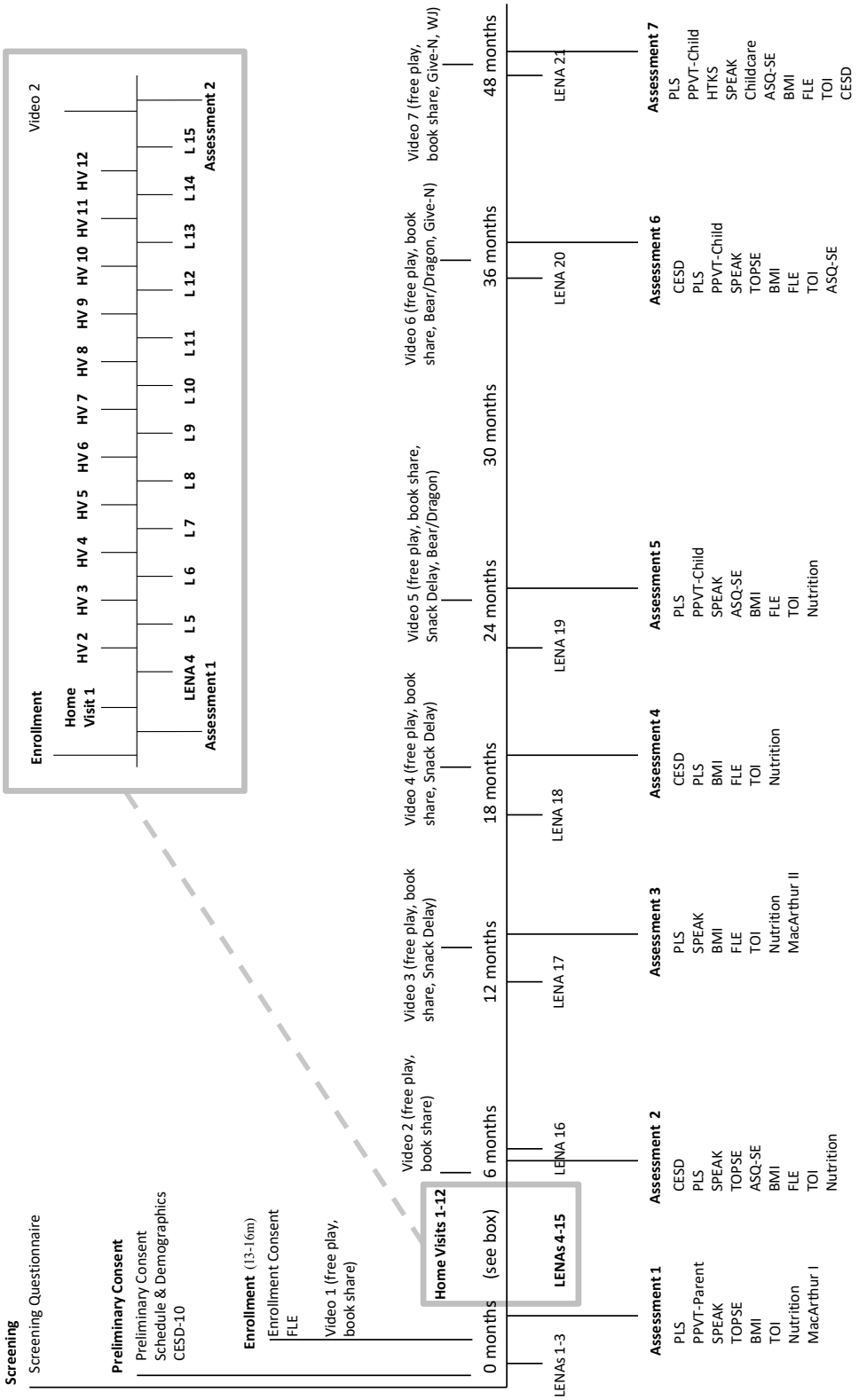
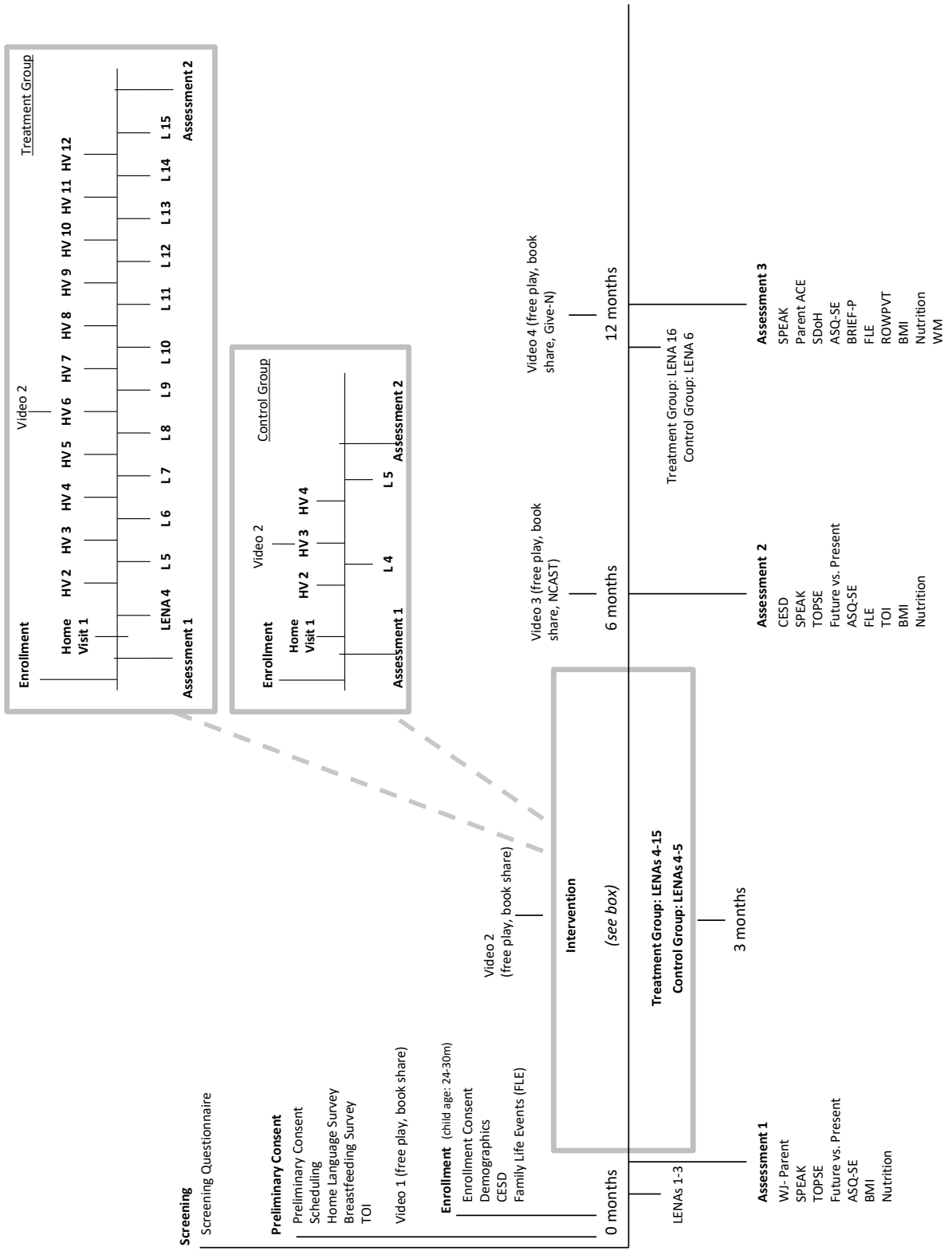


Figure S.2: Timeline Experiment 2



Parental surveys (in alphabetical order):

*ASQ-SE: Ages and Stages Questionnaire – Socioemotional Scale

Breastfeeding Survey: Survey on breastfeeding practices

BRIEF-P: Behavior Rating Inventory of Executive Function – Preschool Version

CESD: Center for Epidemiological Studies - Depression scale

Childcare: Questionnaire about child's early out-of-home care and school experiences

*Demographics: Questionnaire about family's socioeconomic characteristics

FLE: Family Life Events survey of family experiences

Future vs. Present: Survey on parents' time preferences

Home Language Survey: Survey on language(s) used at home

SPEAK: Survey of Parental Knowledge and Expectations about cognitive development

MacArthur I and II: MacArthur-Bates Communicative Development Inventories, short forms levels 1 (8-18 months) and 2 (16-30 months)

Nutrition: Survey designed to measure knowledge of nutrition topics

Parent ACE: Survey on Adverse Childhood Experiences

SDoH: Survey on Social Determinants of Health

TOI: Theories of Intelligence survey

TOPSE: TOol to measure Parenting Self-Efficacy

Direct assessments and recordings (in alphabetical order):

Bear/Dragon: Measure of executive function (inhibition and activation)

BMI: Measure of child's height and weight

***Book Share**: Video recording of a reading activity between caregiver and child using standardized books

***Free Play**: Video recording of a playing activity between caregiver and child using standardized toys

Give-N: Assessment of number word comprehension

HTKS: Head Toes Knees Shoulders measures inhibitory control, working memory, and attention

***LENA**: Audio recordings using the Language ENvironment Analysis technology to measure Conversational Turn Counts (CTCs)

NCAST: Nursing Child Assessment Satellite Training Scale (5-minute long video recording of parents teaching activity to child)

***PLS**: Preschool Language Scale, 5th edition

*__PPVT-Child__: Peabody Picture Vocabulary Test, 4th edition

PPVT-Parent: Peabody Picture Vocabulary Test, 4th edition

*ROWPVT: Receptive One-Word Picture Vocabulary, 4th edition

Snack Delay: Measure of executive function (delay of gratification)

WJ: Woodcock-Johnson tests of cognitive abilities and tests of achievement

*WM: Woodcock-Muñoz tests of expressive vocabulary and oral comprehension

*Measures used for the present analysis. Primary outcomes are indicated in bold font.

Tables S.1 and S.2 below provide summary statistics for each of the assessments used in the analysis as well as for the conversational turn counts (CTCs) measured by LENA. Note that child assessments are presented as raw scores (not standardized) for easier interpretability.

Table S.1: Descriptive statistics CTCs and child outcomes (Exp. 1)

Variable	Mean	SD	Q25	Q50	Q75	N
CTCs recording #1	20.64	11.43	11.82	18.32	28.77	202
CTCs recording #2	19.18	11.47	10.71	17.00	25.14	205
CTCs recording #3	20.03	13.00	11.02	17.50	27.28	200
PLS assessment #1	18.85	1.48	18.00	19.00	19.00	197
PLS assessment #2	23.80	4.65	20.00	23.00	27.00	157
PLS assessment #3	29.01	5.35	24.00	30.00	33.00	152
PLS assessment #4	33.06	5.23	31.00	34.00	36.00	153
PLS assessment #5	36.83	5.60	34.00	37.00	40.00	154
PLS assessment #6	45.03	6.28	41.00	44.00	49.00	106
PPVT assessment #5	37.68	18.80	25.00	37.00	50.00	152
PPVT assessment #6	58.98	21.10	50.00	58.00	72.00	106
ASQ-SE assessment #2	34.90	24.65	15.00	30.00	50.00	157
ASQ-SE assessment #5	47.61	36.94	25.00	40.00	60.00	153
ASQ-SE assessment #6	40.28	28.52	20.00	35.00	55.00	143

Note: Mean, standard deviation, first, second, and third quartile of the CTC recordings and child outcomes. N gives the number of observations.

Table S.2: Descriptive statistics CTCs and child outcomes (Exp. 2)

Variable	Mean	SD	Q25	Q50	Q75	N
CTCs recording #1	27.50	21.92	12.48	22.72	36.81	83
CTCs recording #2	27.09	18.45	14.48	22.01	36.70	82
CTCs recording #3	25.39	19.39	8.67	22.91	39.67	72
ROWPVT English	22.60	15.14	11.00	21.00	35.00	67
ROWPVT Spanish	24.66	14.32	10.00	25.00	37.00	67
WM Analogies English	1.18	2.21	0.00	0.00	2.00	61
WM Analogies Spanish	1.67	2.11	0.00	1.00	3.00	61
WM Picture Vocab English	11.10	6.29	6.00	10.00	16.00	61
WM Picture Vocab Spanish	11.25	5.22	8.00	12.00	15.00	61
ASQ-SE assessment #1	33.36	1.16	32.76	33.06	33.52	82
ASQ-SE assessment #2	35.07	0.84	34.57	35.00	35.36	69
ASQ-SE assessment #3	37.18	0.92	36.57	36.93	37.57	68

Note: Mean, standard deviation, first, second, and third quartile of the CTC recordings and child outcomes. N gives the number of observations.

B Enrollment and randomization

Exp. 1 targeted English-speaking mothers or fathers with a child aged 13 to 16 months at enrollment while Exp. 2 targeted Spanish-speaking mothers with a child aged 24 to 30 months at enrollment. We followed families over four years in Exp. 1 and one year in Exp. 2.

Both studies targeted low-SES families living in the Chicagoland area. Specifically, participants were eligible if their household income was below 200% of the federal poverty line and if their highest degree attained was at or below a Bachelor's degree. Parents also had to be older than 18 years old at the start of the study, have legal custody of their child and live with them, and speak English (Exp. 1) or Spanish (Exp. 2) as their primary home language. Additionally, in the first experiment, children had to hear English at least 80% of the time throughout the week. Children with significant cognitive or physical impairments were not eligible to participate in the study.

The recruitment was done from December 2014 through December 2017 for Exp. 1 and from June 2016 to July 2017 for Exp. 2 using postings at childcare centers, libraries, health clinics, local stores, public transportation, and community organizations serving low-income populations. We aimed for a sample size of 200 participants (100 in the treatment group and 100 in the control group) for Exp. 1 based on power analyses using data from a pilot implementation

of the intervention (Suskind et al. (2016)). For Exp. 2, we targeted a baseline sample size of a minimum of 90 participants. Our final enrolled samples include 206 participants in Exp. 1 and 91 participants in Exp. 2. CONSORT diagrams presenting the number of observations at each stage of the enrollment process and throughout each study are shown below, Figure S.3 and S.4.

In the figures, VS1-VS7 refers to video sessions, while A1-A7 refers to assessments (surveys and child tests). The number of observations at each assessment was calculated based on having responded to at least one survey or having completed at least one child test. During the last time point of Experiment 1 (VS7/A7), due to the start of the COVID-19 pandemic, assessors had to stop visiting families and could not conduct the child assessments and video sessions. As a result, we have child assessment data for less than 30% of the sample and the video sessions were done at home by the parents themselves using their cell phones, which did not provide high-quality enough audio for acoustic processing. We therefore had to exclude that time point from the analysis.

Figure S.3: CONSORT diagram Experiment 1

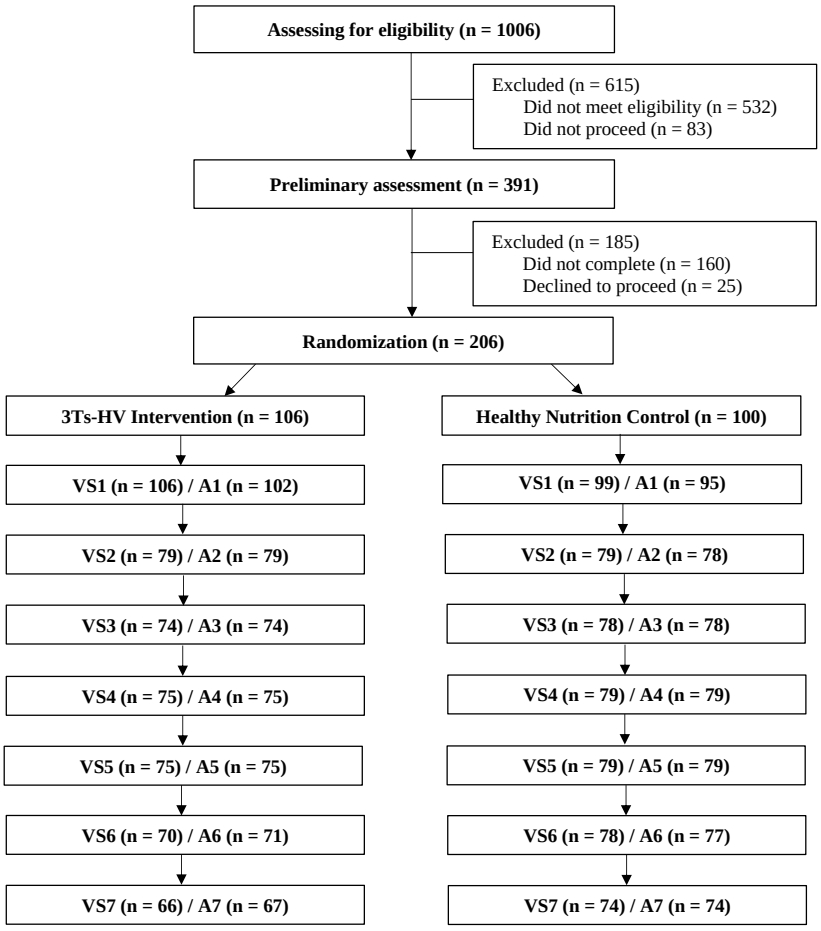
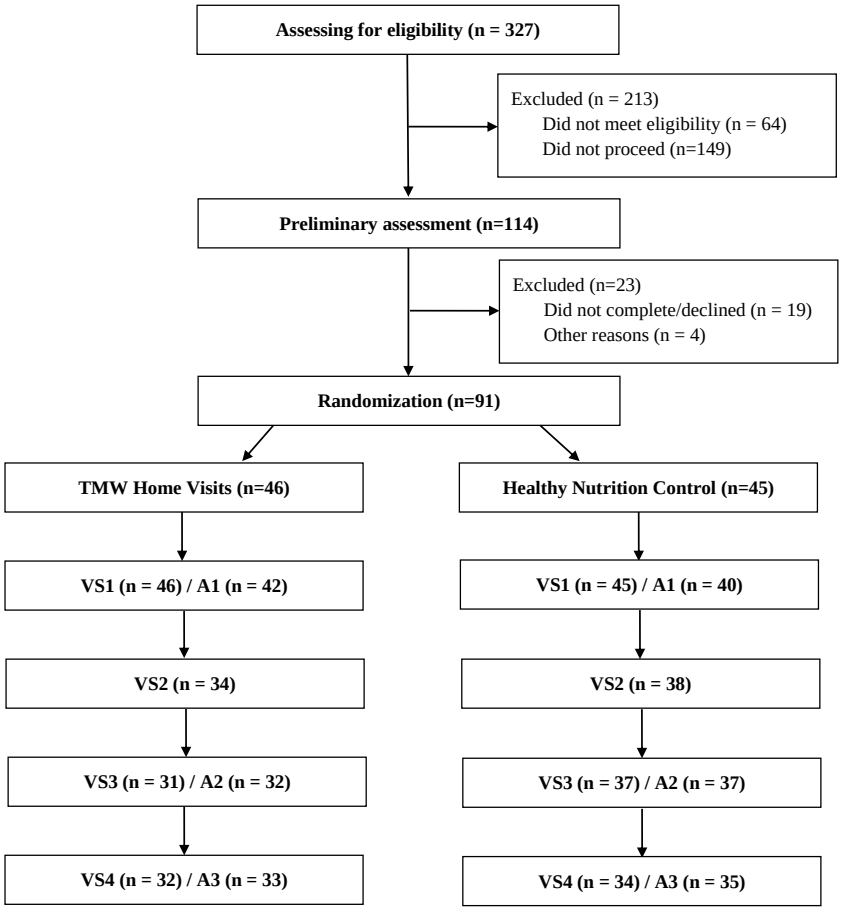


Figure S.4: CONSORT diagram Experiment 2



The two field experiments are registered on clinicaltrials.gov, under the references NCT02216032 (Exp. 1) and NCT03076268 (Exp. 2), where the exact protocols are thoroughly described. In Exp. 1, participants were randomly assigned to either the Treatment or Control condition as part of a matched pair. Pairs were matched by 1) age of the child and 2) averaged conversational turn counts from the 3 baseline LENA recordings. Assignment to Control or Treatment of each member of a pair was decided through a coin flip. In Exp. 2, participants were randomly assigned to either the Treatment or Control condition based on a randomization table. The website Research Randomizer was used to generate a set of 91 unique, unsorted numbers. The resulting numbers were entered into an Excel spreadsheet where the 91 unique, unsorted numbers were sorted into the two conditions. In both experiments, the randomization, enrollment, and

assignment were performed by the research team at the TMW Center.

In Tables S.3 and S.4 below, we provide summary statistics on the socioeconomic characteristics of our sample for each study and check that the treatment and control groups are balanced on those characteristics at baseline and throughout the study. Several important features of our sample can be noted. First, even though there were no differences in recruitment efforts to enroll mothers rather than fathers, almost all of our participants are mothers (97%) in Exp. 1. In Exp. 2, all participants are mothers per enrollment criteria. Participants are on average 29 and 33 years old at enrollment in Exp. 1 and Exp. 2., respectively. About 25% of our subjects are married or partnered in Exp. 1 compared to approximately 70% in Exp. 2.

The vast majority of parents identify as Black or African American in Exp. 1 and all identify as Hispanic or Latino in Exp. 2. Examining education levels, while both studies targeted participants without a college education, our first sample is on average more educated with only 30% of participants having a high-school diploma or below as their highest degree compared to 70% in our second study. Roughly half of the participants are employed at baseline in both studies, and, consistent with our eligibility criteria targeting low-income families, a large share of participants receive benefits such as the Special Supplemental Nutrition Program for Women, Infants, and Children (WIC) and cash assistance or food stamps (LINK card), with a higher share in Exp. 1.

Focusing on the differences between treatment and control groups (columns T-C), Tables S.3 and S.4 show that the two groups are on average comparable at baseline and at later time points although we have some attrition throughout the studies (see CONSORT diagrams above). In Exp. 1, only the difference in the probability of having a LINK card approaches standard significance levels (p-values around 10-20%), with the treatment group being 10 percentage points less likely to have a card. The treatment and control groups in Exp. 2 are significantly different in terms of the probability of being employed and of receiving WIC benefits, with the treatment group being about 30 percentage points less likely to be employed, and 13 percentage points less likely to receive WIC benefits. To account for those differences in the analysis, we explore various robustness tests including controlling for individual fixed effects that capture any time-invariant differences between participants, including the differences noted above.

Table S.3: Balancing checks at each time point (Exp. 1)

	VS1		VS2		VS3		VS4		VS5		VS6	
Variable	MC	T-C	MC	T-C	MC	T-C	MC	T-C	MC	T-C	MC	T-C
Parent = female	0.97	0.00	0.97	0.01	0.97	0.01	0.97	0.01	0.97	0.01	0.97	0.01
	(0.17)	*0.92*	(0.16)	*0.56*	(0.16)	*0.57*	(0.16)	*0.58*	(0.16)	*0.58*	(0.16)	*0.61*
Child = female	0.51	-0.07	0.48	-0.01	0.50	-0.04	0.51	-0.04	0.51	-0.04	0.50	-0.06
	(0.50)	*0.31*	(0.50)	*0.87*	(0.50)	*0.62*	(0.50)	*0.63*	(0.50)	*0.63*	(0.50)	*0.49*
Age parent	28.67	0.61	29.20	-0.10	29.25	0.25	29.18	0.24	29.18	0.27	29.26	0.67
	(6.96)	*0.52*	(7.41)	*0.93*	(7.47)	*0.83*	(7.36)	*0.83*	(7.36)	*0.81*	(7.40)	*0.56*
Married/partnered	0.25	0.02	0.25	0.05	0.26	0.06	0.25	0.07	0.25	0.05	0.26	0.07
	(0.44)	*0.73*	(0.44)	*0.48*	(0.44)	*0.41*	(0.44)	*0.36*	(0.44)	*0.46*	(0.44)	*0.34*
Black	0.83	-0.05	0.83	-0.05	0.81	-0.01	0.82	-0.03	0.82	-0.02	0.82	-0.02
	(0.37)	*0.37*	(0.38)	*0.47*	(0.40)	*0.84*	(0.39)	*0.63*	(0.39)	*0.78*	(0.39)	*0.81*
White	0.08	-0.04	0.08	-0.03	0.08	-0.03	0.08	-0.02	0.08	-0.02	0.08	-0.04
	(0.28)	*0.30*	(0.27)	*0.49*	(0.28)	*0.52*	(0.27)	*0.54*	(0.27)	*0.54*	(0.27)	*0.36*
Hispanic/Latino	0.12	0.04	0.13	0.04	0.12	0.05	0.13	0.03	0.13	0.02	0.13	0.01
	(0.33)	*0.42*	(0.33)	*0.50*	(0.33)	*0.44*	(0.33)	*0.56*	(0.33)	*0.72*	(0.34)	*0.80*
Uses English only w/ child	0.96	-0.04	0.96	-0.06	0.96	-0.06	0.96	-0.06	0.96	-0.04	0.96	-0.05
	(0.20)	*0.19*	(0.19)	*0.12*	(0.20)	*0.19*	(0.19)	*0.17*	(0.19)	*0.28*	(0.20)	*0.25*
Uses Eng. and Span. w/ child	0.04	0.03	0.04	0.04	0.04	0.03	0.04	0.03	0.04	0.02	0.04	0.02
	(0.20)	*0.42*	(0.19)	*0.31*	(0.20)	*0.46*	(0.19)	*0.44*	(0.19)	*0.66*	(0.20)	*0.61*
Educ $\leq$ HS	0.30	0.05	0.30	-0.00	0.30	0.01	0.30	0.00	0.30	0.00	0.28	0.03
	(0.46)	*0.48*	(0.46)	*1.00*	(0.46)	*0.91*	(0.46)	*0.97*	(0.46)	*0.97*	(0.45)	*0.67*
Employed	0.52	-0.03	0.48	-0.01	0.46	0.03	0.47	0.01	0.47	0.01	0.46	0.02
	(0.50)	*0.63*	(0.50)	*0.87*	(0.50)	*0.75*	(0.50)	*0.89*	(0.50)	*0.89*	(0.50)	*0.77*
Receives WIC	0.75	0.01	0.76	-0.03	0.76	0.01	0.76	-0.01	0.76	-0.01	0.76	-0.01
	(0.44)	*0.91*	(0.43)	*0.72*	(0.43)	*0.92*	(0.43)	*0.85*	(0.43)	*0.85*	(0.43)	*0.85*
Has LINK card	0.84	-0.09	0.80	-0.09	0.80	-0.10	0.80	-0.09	0.80	-0.09	0.79	-0.11
	(0.37)	*0.10*	(0.40)	*0.20*	(0.40)	*0.16*	(0.40)	*0.20*	(0.40)	*0.20*	(0.41)	*0.13*

Note: MC gives the mean of the variable in the control group, with the standard deviation in parentheses below. T-C gives the difference between the treatment and control groups, with the p-value in italics below. The sample used at each time point is determined by the number of participants in each video session (see CONSORT diagrams in the Supplementary Materials).

Table S.4: Balancing checks at each time point (Exp. 2)

	VS1		VS2		VS3		VS4	
Variable	MC	T-C	MC	T-C	MC	T-C	MC	T-C
Parent = female	1.00	0.00	1.00	0.00	1.00	0.00	1.00	0.00
	(0.00)	.	(0.00)	.	(0.00)	.	(0.00)	.
Child = female	0.44	-0.12	0.45	-0.12	0.43	-0.08	0.44	-0.10
	(0.50)	*0.25*	(0.50)	*0.29*	(0.50)	*0.52*	(0.50)	*0.43*
Age parent	33.00	0.48	32.96	1.64	33.45	1.51	33.69	0.99
	(8.83)	*0.77*	(9.25)	*0.37*	(8.94)	*0.41*	(8.83)	*0.59*
Married/partnered	0.69	0.14	0.74	0.17	0.73	0.14	0.71	0.14
	(0.47)	*0.13*	(0.45)	*0.05*	(0.45)	*0.15*	(0.46)	*0.18*
Black	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	(0.00)	.	(0.00)	.	(0.00)	.	(0.00)	.
White	0.29	0.10	0.32	0.01	0.30	0.03	0.29	0.02
	(0.46)	*0.31*	(0.47)	*0.94*	(0.46)	*0.83*	(0.46)	*0.87*
Hispanic/Latino	1.00	0.00	1.00	0.00	1.00	0.00	1.00	0.00
	(0.00)	.	(0.00)	.	(0.00)	.	(0.00)	.
Uses English only w/ child	0.02	0.02	0.00	0.06	0.00	0.06	0.00	0.06
	(0.15)	*0.57*	(0.00)	*0.16*	(0.00)	*0.15*	(0.00)	*0.16*
Uses Spanish only w/ child	0.64	-0.04	0.66	-0.04	0.65	-0.00	0.62	0.01
	(0.48)	*0.73*	(0.48)	*0.73*	(0.48)	*0.98*	(0.49)	*0.95*
Uses Eng. and Span. w/ child	0.33	0.01	0.34	-0.02	0.35	-0.06	0.38	-0.07
	(0.48)	*0.89*	(0.48)	*0.87*	(0.48)	*0.60*	(0.49)	*0.56*
Educ $\leq$ HS	0.69	0.12	0.74	0.06	0.73	0.08	0.71	0.08
	(0.47)	*0.21*	(0.45)	*0.57*	(0.45)	*0.46*	(0.46)	*0.49*
Employed	0.51	-0.32	0.50	-0.29	0.46	-0.33	0.44	-0.32
	(0.51)	*0.00*	(0.51)	*0.01*	(0.51)	*0.00*	(0.50)	*0.00*
Receives WIC	0.64	-0.12	0.66	-0.13	0.68	-0.13	0.65	-0.08
	(0.48)	*0.24*	(0.48)	*0.27*	(0.47)	*0.29*	(0.49)	*0.49*
Has LINK card	0.62	-0.06	0.61	-0.05	0.62	-0.07	0.59	-0.03
	(0.49)	*0.58*	(0.50)	*0.70*	(0.49)	*0.55*	(0.50)	*0.84*

Note: MC gives the mean of the variable in the control group, with the standard deviation in parentheses below. T-C gives the difference between the treatment and control groups, with the p-value in italics below. The sample used at each time point is determined by the number of participants in each video session (see CONSORT diagrams in the Supplementary Materials).

C Multi-step algorithm

C.1 Speaker Labeling

Building a reliable classifier to label the speakers in our parent-child interaction data is the first and most critical part of the model. To do so, we used a top-performing audio classification neural network architecture, Pretrained Audio Neural Networks (PANN) (Kong et al. (2020)), and adapted it to our data to maximize accuracy in our specific framework.

Adapting the PANN classifier

The PANN model was trained on large-scale audio datasets and outperforms most of the existing models including Convolutional Neural Networks (CNNs), Gaussian Mixture Models (GMMs), and Support Vector Machines (SVMs) models (Kong et al. (2020)). The model is highly robust and can be transferred into other audio pattern recognition tasks as well. PANN is built by applying Short Time Fourier Transforms (STFTs) to capture frequency information as inputs to the model. The model contains neural networks having convolutional layers with batch normalization and the Rectified Linear Unit (ReLU) functions to capture nonlinear behaviors in the audio. In addition, various tuning techniques such as data balancing, data augmentation, adjusting hop sizes, adjusting embedding dimensions, and trying different sample rates and CNN layers were applied to find the optimal model configuration.

The algorithm was trained on 1.9 million audio clips and is able to detect 527 sound classes including animal, musical, and human speech (Kong et al. (2020)). Among various sound classes, human speech has one of the highest Average Precision (AP), above 80% (see Figure 3 in Kong et al. (2020)). The algorithm is language agnostic, but it should be noted that this does not guarantee that its performance on languages it was not trained on is as high. For our study, we tuned PANN to make it more aligned with our specific research questions. In particular, our goal was to use PANN's computational capacity more efficiently since we only have a few classes of interest. Therefore, the last layer of PANN was retrained using pre-labeled datasets from YouTube (CHILDES Dataset, MacWhinney (2000)) and HomeBank (VanDam et al. (2016)) to have a multiclass prediction for the three classes that are relevant to our analysis: "mother", "father", and "child".[1]

[1]Specifically, we used 204 minutes total from the VanDam datasets and 44 minutes total from the CHILDES dataset. Additional information on those two datasets can be found at https://talkbank.org/homebank/access/Public/VanDam-5minute.html and https://talkbank.org/childes/topics/youtube.html, respectively.

Specifically, we used a learning rate of 0.001, cross-entropy loss, five epochs, Adam optimizer, and PyTorch. 80% of the data was used for training and 20% for testing.

Maximizing precision in our data

After adapting the PANN classifier to predict those three classes of interest, we applied it to our video session data to assess its performance. To do so, we first combined the re-trained classifier with Google's WebRTC Voice Activity Detection (VAD) model.[2] This model predicts the probability of human voice in each audio segment and allows us to screen out silence and non-human sounds from the recordings. To determine the probability thresholds that maximize accuracy, we trained a decision tree classifier that uses the four probabilities (human voice, female, male, child) as inputs and the labels from our transcribers as outcomes.

The transcriptions of the video recordings come from Exp. 1. All recordings were transcribed and were subjected to various validation and reliability checks and we therefore use them as the ground truth to train the decision tree. In this step, we had to overcome the challenges of imbalanced data due to low child speech compared to adult speech and non-speech segments. To compensate for this class imbalance, we opted for a weighted decision tree using the under-sampling method. We performed a grid search to find the optimal class weights for child, adult, and blank and chose class weights that maximize the precision of the female label (97% of our adult participants being females) and minimize the risk of predicting child speech as female. We also leveraged the variation in the age of the children in our study to train separate decision trees for different age groups.

Table S.5 reveals that our model achieves precision scores between 0.64 and 0.72 for our class of interest (female speaker), depending on the age group (i.e., video sessions 1 to 6), when compared to our human transcripts. This is 7 to 20% higher than the performance of the LENA technology, which is widely used in research on parent-child interactions (Cristia et al. (2021)).

[2]https://github.com/wiseman/py-webrtcvad

Table S.5: Precision of the female prediction by age of the child

Age group	Precision	N
13-16 months (VS1)	0.72	2217
19-22 months (VS2)	0.68	9376
25-28 months (VS3)	0.70	16979
31-34 months (VS4)	0.64	24315
37-40 months (VS5)	0.70	31054
49-52 months (VS6)	0.66	36563

Note: Precision = True Positive / (True Positive + False Positive). N corresponds to the number of seconds with a speaker label in our recordings. Precision is calculated based on the second-level data.

Next, Figure S.5 displays the confusion matrices that provide the complete comparison between our model's predictions and the labels coming from our human transcriptions. The matrices reveal that a relatively small share of child speech is predicted as female speech (between 2.8% and 5.7% - last column, second row), which was another potential threat to the validity of our analyses, as we wanted to avoid confounding the characteristics of children's voice with the characteristics of the parents' voice.

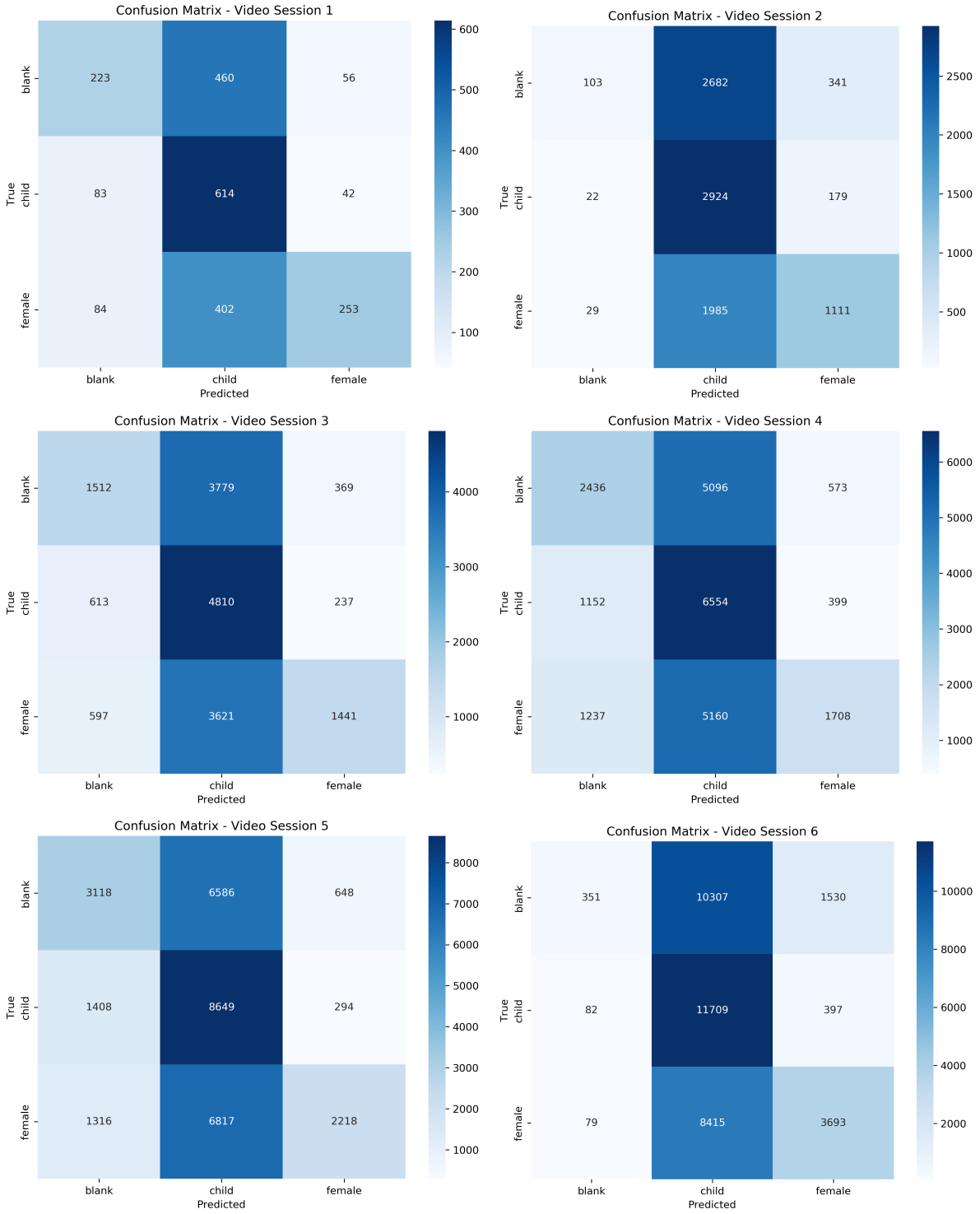


Figure S.5: Confusion matrices between our predictions (x-axis) and human annotations (y-axis). The number in each cell corresponds to the number of frames in the intersection.

Finally, we matched the age-specific decision trees from Exp. 1 to the audio recordings of Exp. 2, for which we do not have validated and timestamped transcriptions allowing us to conduct the same precision analysis. Specifically, age group 25-28 months in Exp. 1 was matched with

age group 24-30 months in Exp. 2, age group 31-34 months was matched with age groups 27-33 and 30-36 months in Exp. 2, and age group 37-40 months was matched with age group 36-42 (see Section A of the Supplementary Materials for the timeline of the two studies and age of the children at each video session).

C.2 Feature Extraction

The second part of our acoustic model consists of generating variables that characterize the voice of the parent. To do so, we took an agnostic and data-driven approach and extracted a large range of frequency features commonly used in speech analysis. Features were chosen to encompass various potential aspects of parental speech. In particular, we borrowed from research on emotional responses to music (e.g., Er and Aydilek (2019), Weninger et al. (2013)) or to sound effects used in games or films (e.g., Sin et al. (2019)), or research on voice features used to detect clinical depression (e.g., Low et al. (2010), Taguchi et al. (2018), Rejaibi et al. (2022)).

Following the features used in those studies, we generated a total of 171 frequency variables capturing Chroma (12 features), Contrast (7 features), Mel-Frequency Cepstral Coefficients (MFCC) (20 features), log-scaled Mel-spectrograms (128 features), Spectral bandwidth (1 feature), Centroid (1 feature), Flatness (1 feature), and F0 (1 feature), using the Python package called *librosa* (McFee et al. (2015)). Note that the extraction of those features is agnostic to the language of the speakers (or the content of the speech more generally), it only captures the acoustic characteristics of the voice. We extracted the features using a sample rate of 16000Hz. Each audio file was windowed into one-second-long non-overlapping windows. For each window, we used a Fast Fourier Transform (FFT) size of 2048, aligned the frame length on the FFT size, and used a hop length of 512. For each feature, we took the mean value over the window. We then moved to the next window. This approach allowed us to achieve a high resolution in the frequency space while keeping the model tractable from a computational perspective. The same extraction parameters were used for all features except F0, which is not based on Short-Term Fourier Transform (STFT). For F0, we used a minimum frequency of 65Hz and a maximum frequency of 523Hz (to cover the typical range of human voice frequency), a frame and hop lengths of 480 and 120, respectively.[3] We then matched features with speaker labels in our audio files at the second level.

[3]More detailed explanations can be found in Mauch and Dixon (2014).

Chroma, our first array of features, is a mid-level representation of sound. Known for their robustness across different sound sources, Chroma features encapsulate characteristics such as pitch and beats in music, while higher-level representations encompass mood and style (Bello). In our study, Chroma-based features represent the short-time energy of the signal. They range from 0 to 11 and capture different musical notes, respectively, C, C#, D, D#, E, F, F#, G, G#, A, A#, and B, with spikes in the chromagram indicating which musical notes are more dominant in audio. Mel-spectrograms are derived through STFT. These features aim to evenly distribute frequencies into bins perceived as equidistant by humans. Our study employs a scale of 128 bins, ranging from 0 to 127. Notably, the bins most acutely detectable by human ears are those in the middle range, to which the human auditory system is most sensitive. We also extracted MFCC features, that reflect the overall spectral envelope of speech and are highly complementary with the Mel-spectrogram; they can be interpreted as a compressed version of it. MFFC features are commonly used for a wide range of speech processing tasks, such as automatic speech recognition and speech synthesis. These acoustic features do not have a direct interpretation, but some examples of uses include predicting speakers' identity and gender, capturing prosodic and other paralinguistic information, or classifying distinct emotional categories such as sadness, anger, and joy. In our analysis, MFCC features comprise 20 distinct values. Another type of features employed in our study are contrast features, which denote the decibel difference between the top and bottom of a spectrum. This metric reflects the level of energy variation in a sound, with low-contrast sounds typically exhibiting more noise and less clarity. More specifically, spectral contrast features capture the difference between spectral peaks and valleys within sub-bands of an audio signal. These features are commonly used in speech and audio processing to distinguish different sound sources and can help highlight the clarity of speech by emphasizing differences between harmonic (speech-like) components and noise (for an example of application, see Nogueira et al. (2016)). In our analysis, contrast features comprise 7 distinct bands that capture the average difference between the maximal and minimal energy value for each frequency in that band. A higher contrast indicates more signal in that frequency band, which implies that a higher value represents a clearer signal, while a lower value suggests that the frequency band comprises more noise. Lastly, spectral bandwidth, another feature extracted in our study, can be interpreted as the standard deviation of the spectrogram. This feature contributes to our understanding of the frequency distribution within the audio. Spectral centroid is the center of mass of the spectrogram, indicating the most dominant frequency in the audio, and spectral flatness measures the amount of noise in the audio (for instance, white noise has a high spectral flatness, while a noiseless audio has low spectral flatness). Overall, this large set of features

covers a variety of sound characteristics that allow us to explore multiple acoustic aspects of parental speech.

C.3 Construction of the indices

After isolating parents' speech inputs and acoustically characterizing them, the goal is to identify the features that best predict cognitive and non-cognitive skills. The identification of such features relies on two main steps. First, we calculate the mean and standard deviation of each feature, for each participant, and for each video session, from the distribution at the second level. This approach is commonly used in acoustic research to model time and frequency domain features (e.g., Whittingslow et al. (2020) or Krajewski et al. (2009) for speech processing specifically) and Zhang et al. (2020) show that learning from such aggregated statistics achieves similar mean squared error (MSE) as from direct observations. This aggregation step allows us to reduce the dimension of our data to two values per feature (i.e. 342 values), per participant, and time point, and match our child assessments, which also vary at the participant and time point level.

With these metrics in hand, the next step is to identify which of the 342 speech features holds the most import for children's skill formation. To do so, we use a Bayesian variable-selection procedure and apply it to our control group data. The rationale behind that sample restriction is to avoid capturing the effect of the intervention, which could simultaneously affect parents' speech and child outcomes, and create a spurious correlation. The Bayesian procedure relies on Stochastic Search Variable Selection (SSVS) (see Bainter et al. (2020)). It consists of estimating the following model:

$$y_i = \beta_0 + \delta_1\beta_1 X1_i + \delta_2\beta_2 X2_i + ... + \delta_p\beta_p Xp_i + \epsilon_i$$

where $\delta_j \in \{0, 1\}$ indicates whether predictor Xj (a speech feature, in our case) is included in the model or not, β_j is the coefficient associated with predictor Xj, and y is the child outcome measure, assessed at the last time point (i.e., assessment 6 for Exp. 1 and assessment 4 for Exp. 2, see Section A of the Supplementary Materials).

For each δ_j, we started with a probability of 0.5, meaning that each feature has a 50% chance of being included in the model. For β_j, we provided a prior of $Uniform(-\infty, +\infty)$, that is to say, we are agnostic about the parameters' values. The two types of children's skills we

focus on are language skills and socioemotional skills. The language score is a z-score of our different language assessments (PLS and PPVT in Exp. 1; ROWPVT and Woodcock-Muñoz tests in Exp. 2, see Section A of the Supplementary Materials) while the socioemotional score corresponds to the ASQ-SE in both experiments. For each score, in each study, we used Markov Chain Monte Carlo (MCMC) to sample 10,000 models and calculate the fraction of appearance ($\delta_j > 0$) of each predictor as well as its correlation with the score (β_j).

To mitigate the risk of overfitting, we implemented for each score and each study a five-fold cross-validation where each fold was randomly divided between a training sample (80% of the data in the fold) and a test sample (the remaining 20% of the data). In each fold, we used the training sample to estimate the β and δ coefficients and used those estimates to predict the index in the remaining 20% of the sample. Since each participant is in the test sample of exactly one fold and, by construction, excluded from the training sample of that same fold, the cross-validation procedure allows us to cover the entire sample (5x20%) while avoiding information leakage.

Our speech indices therefore correspond to the following predictions:

$$\hat{y_i} = \hat{\beta_0} + \hat{\delta_1}\hat{\beta_1}X1_i + \hat{\delta_2}\hat{\beta_2}X2_i + ... + \hat{\delta_p}\hat{\beta_p}Xp_i$$

where the vectors $\hat{\beta} = (\hat{\beta_0}, \hat{\beta_1}, \hat{\beta_2}, ..., \hat{\beta_p})$ and $\hat{\delta} = (1, \hat{\delta_1}, \hat{\delta_2}, ..., \hat{\delta_p})$ can vary across folds. In Table S.6 below, we analyze how similar the vectors $\hat{\beta} \odot \hat{\delta}$ (element-wise products of vectors $\hat{\beta}$ and $\hat{\delta}$) are across folds by doing systematic pairwise comparisons. In general, the different folds yield similar vectors, with cosine values mostly around or above 0.5.

For the treatment group, which was excluded from the SSVS analysis to avoid confounding the correlation between the features and child outcomes with the effect of the intervention, we took the mean of $\hat{\beta} \odot \hat{\delta}$ across the five folds. We also imputed that mean to the control group participants who had missing indices due to missing scores (y variable, missing for 8% of the control data).

This procedure results in each participant in our dataset having two predicted indices (one for language skills and one for socioemotional skills) corresponding to each video session they participated in, based on the full set of features extracted from the audio data. While the values those indices take are not directly interpretable in terms of specific speaking styles, they are driven by the features that are most predictive of children's language and socioemotional skills.

Table S.6: Cosine similarity of the SSVS coefficients across folds for each index

		Index Language Skills					Index Socioemo. Skills				
	Fold	1	2	3	4	5	1	2	3	4	5
Experiment 1	1	1.00					1.00				
	2	0.98	1.00				0.61	1.00			
	3	0.98	0.97	1.00			0.65	0.28	1.00		
	4	0.98	0.98	0.97	1.00		0.70	0.71	0.45	1.00	
	5	0.96	0.96	0.96	0.97	1.00	0.49	0.73	0.28	0.76	1.00
Experiment 2	1	1.00					1.00				
	2	0.43	1.00				0.79	1.00			
	3	0.41	0.62	1.00			0.68	0.70	1.00		
	4	0.55	0.71	0.59	1.00		0.47	0.37	0.42	1.00	
	5	0.60	0.59	0.61	0.73	1.00	0.73	0.73	0.66	0.45	1.00

Note: Cosine similarity between the vectors of coefficients ($\delta \odot \beta$) predicted by the SSVS model in each fold of the cross-validation procedure. Values range from -1 to 1, with -1 representing two opposite vectors, 0 representing two orthogonal vectors, and 1 representing two proportional vectors. The top part of the table provides the cosine values for the vectors comparison for the two types of indices in the first experiment, while the bottom part of the table provides the cosine values for the two indices in the second experiment.

C.4 Top predictors of child outcomes for each index

Table S.7 below presents for each experiment and each index the top 30 features from the SSVS analysis. The features are ranked along their marginal inclusion probability (MIP), which is calculated as the average MIP across the five folds of the cross-validation. The MIPs are captured by the $\hat{\delta}$'s in the model described in Section C.3 and can be interpreted as weights (although not standardized to sum to one) in the calculation of the indices.

Overall, the MIP distributions suggest that a small set of features drive the indices, with the top five to ten features having relatively high weights in all cases. MFCCs capturing the spectral envelope of parents' speech are among the strongest predictors of children's language and socioemotional outcomes in both experiments, but notably dominate the language index in Exp. 1. Closely associated with those features, Mel-spectograms (Mels, in the table) are also among the main drivers of the two indices in both experiments, with middle to high-range bins representing at least half of the top-30 feature set. Other important predictors include contrast features (Contr, in the table), which, as explained in Section C.2, can be interpreted as capturing the clarity of parental speech, and have particularly high weights in the socioemotional index in the second experiment (Contrast 5, 3, and 4 having respective MIPs of 0.77, 0.39, and 0.26), but are

also strong predictors of language skills in both experiments (Contrast 4 has an MIP of 0.51 in Exp. 1, Contrast 5 has an MIP of 0.37 in Exp. 2). Chromatic features of speech (Chrom, in the table) reflecting different musical notes (see section C.2) also appear among the top predictors of all indices. Last, one can note that the standard deviation of the spectrogram, captured by the spectral bandwidth (Spec bw, in the table) is an important driver of the language index in the second experiment.

Table S.7: Top 30 features of each index

Exp. 1				Exp. 2			
Language index		Socioemo. index		Language index		Socioemo. index	
Feat.	MIP	Feat.	MIP	Feat.	MIP	Feat.	MIP
Mfcc0 (mn)	0.70	Mfcc10 (sd)	0.62	Spec bw (sd)	0.60	Contr5 (sd)	0.77
Mfcc6 (mn)	0.54	Mels121 (mn)	0.59	Chrom8 (mn)	0.47	Contr3 (sd)	0.39
Contr4 (mn)	0.51	Mfcc7 (sd)	0.57	Mels3 (sd)	0.41	Mfcc6 (mn)	0.35
Mfcc4 (mn)	0.46	Mels123 (mn)	0.48	Contr5 (sd)	0.37	Mfcc0 (mn)	0.29
Mels125 (mn)	0.45	Mels119 (mn)	0.37	Mels63 (mn)	0.36	Contr4 (sd)	0.26
Chrom11 (sd)	0.40	Mfcc7 (mn)	0.37	Mels84 (mn)	0.35	Mels116 (mn)	0.24
Centroid (mn)	0.40	Mels110 (mn)	0.33	Mels16 (mn)	0.33	Mfcc3 (sd)	0.23
Mels93 (mn)	0.40	Mels73 (mn)	0.32	Mels82 (sd)	0.27	Mels120 (sd)	0.21
Mfcc3 (sd)	0.40	Mels88 (sd)	0.32	Centroid (sd)	0.23	Chrom2 (mn)	0.20
Mfcc17 (sd)	0.37	Chrom11 (sd)	0.30	Mfcc1 (mn)	0.23	Mfcc1 (mn)	0.19
Mels105 (mn)	0.36	Mels121 (sd)	0.29	Mels66 (mn)	0.22	Mfcc13 (sd)	0.19
Centroid (sd)	0.35	Mels73 (sd)	0.29	Mels4 (sd)	0.21	Mels115 (mn)	0.18
Mels74 (mn)	0.35	Mels46 (mn)	0.29	Mels82 (mn)	0.20	Mels121 (sd)	0.18
Mels46 (mn)	0.35	Mels85 (mn)	0.28	Chrom8 (sd)	0.20	Contr2 (mn)	0.18
Mfcc2 (sd)	0.35	Mels88 (mn)	0.25	Mels16 (sd)	0.20	Mels127 (mn)	0.18
Mels117 (mn)	0.35	Mels110 (sd)	0.24	Mels63 (sd)	0.20	Spec bw (sd)	0.18
Mfcc1 (mn)	0.35	Mels123 (sd)	0.24	Contr2 (sd)	0.19	Mels96 (mn)	0.17
Mels46 (sd)	0.34	Mfcc17 (sd)	0.20	Contr3 (sd)	0.18	Chrom1 (mn)	0.16
Mels94 (mn)	0.34	Mels47 (mn)	0.19	Mels84 (sd)	0.18	Mfcc14 (sd)	0.16
Mels119 (mn)	0.33	Mfcc4 (sd)	0.19	Mfcc7 (sd)	0.17	Mels110 (mn)	0.16
Mfcc7 (mn)	0.32	Mels84 (mn)	0.19	Mels85 (mn)	0.17	Mfcc17 (mn)	0.16
Flatness (mn)	0.32	Mels49 (mn)	0.19	Mels83 (mn)	0.17	Mels75 (mn)	0.16
Chrom8 (mn)	0.30	Mels46 (sd)	0.18	Mels118 (mn)	0.16	Mels9 (mn)	0.15
Mels69 (sd)	0.30	Mels101 (mn)	0.17	Mels12 (sd)	0.15	Mels125 (sd)	0.15
Mels69 (mn)	0.29	Chrom3 (mn)	0.17	Mels72 (sd)	0.15	Mels113 (mn)	0.15
Mfcc9 (sd)	0.29	Mfcc11 (sd)	0.17	Contr4 (sd)	0.15	Mfcc18 (sd)	0.15
Mels68 (mn)	0.28	Mels78 (sd)	0.16	Mels110 (mn)	0.15	Mels69 (mn)	0.15
Mels105 (sd)	0.28	Mels49 (sd)	0.16	Mfcc11 (sd)	0.15	Mels103 (mn)	0.14
Chrom7 (mn)	0.28	Mfcc6 (mn)	0.14	Mels78 (mn)	0.15	Mels86 (mn)	0.14
Mels95 (mn)	0.26	Mels81 (mn)	0.14	Mels66 (sd)	0.14	Mels66 (mn)	0.14

Note: Top 30 features ranked by their marginal inclusion probability (MIP) averaged across the five folds of the SSVS analysis for each index and each experiment. Chrom stands for Chroma, Contr for Contrast, Mfcc for Mel-Frequency Cepstral Coefficients, and Mels for Mel-spectrograms (see section C.2).

C.5 Correlation between the indices and child outcomes

Table S.8 presents the correlation between each skill (language, socioemotional) and the index built to predict it. We do this separately by group and by experiment. Note that we measure skills at the last time point (consistent with the SSVS estimation) and indices are evaluated at each time point. Both are in standard deviation units. Two patterns emerge from the data. First, as expected, given that SSVS was estimated on control-group data, correlations are stronger in the control group. For example, in Exp. 1, control-group correlations cluster around 0.3-0.4 for both indices. The treatment group shows much more variation across time points and substantially larger standard errors, especially for socioemotional skills. In addition, for Exp. 2, control-group correlations are higher on average (typically 0.6-0.7) while treatment-group correlations are attenuated. Indeed, for socioemotional skills, they are negative at most time points (though imprecise). A second pattern from the data is the divergence between groups. We view this as informative rather than a validation failure. This is because the indices identify acoustic features that predict child outcomes in the absence of intervention. In untreated families, those features plausibly co-vary with latent parenting quality and other unobserved drivers of child development. The intervention pushes parents toward certain feature patterns regardless of underlying baseline behavior. This behavioral change mechanically attenuates the correlation between the indices and child outcomes among the treated. That is, the features become a less informative proxy for the latent factors they tracked in the control sample. This is consistent with the program changing not just the level of speech features but also what those features measure. More broadly, it illustrates a Stage-3 external-validity question (SANS framework, List (2020)): a predictive mapping estimated under one set of conditions need not transfer to a population whose underlying behavior has been intentionally shifted. We therefore caution against interpreting the indices as a fixed measurement instrument. Rather, they are best understood as a data-driven characterization of speech features predictive of child outcomes in untreated, low-SES families in our two samples, whose malleability to a parenting intervention is the object of study.

Table S.8: Correlation between indices and child outcomes in each group

	Exp. 1				Exp. 2			
	Language skills		Socioemo. skills		Language skills		Socioemo. skills	
	T	C	T	C	T	C	T	C
t = 1	0.00	0.46	0.17	0.37	-0.22	0.17	-0.24	0.70
	(0.05)	(0.18)	(0.51)	(0.26)	(0.25)	(0.27)	(0.15)	(0.21)
t = 2	0.31	0.31	-0.15	0.40	0.12	0.68	-0.19	1.03
	(0.20)	(0.22)	(0.31)	(0.20)	(0.12)	(0.17)	(0.14)	(0.26)
t = 3	0.21	0.30	-0.52	0.39	0.23	0.59	-0.20	0.54
	(0.22)	(0.25)	(0.40)	(0.33)	(0.19)	(0.16)	(0.15)	(0.18)
t = 4	0.35	0.24	0.22	0.26	-0.19	0.58	-0.05	0.66
	(0.19)	(0.19)	(0.39)	(0.25)	(0.22)	(0.18)	(0.19)	(0.29)
t = 5	0.09	0.18	-0.60	0.16	-	-	-	-
	(0.22)	(0.16)	(0.40)	(0.18)	-	-	-	-
t = 6	0.55	0.43	0.10	-0.34	-	-	-	-
	(0.27)	(0.19)	(0.38)	(0.24)	-	-	-	-

Note: OLS regressions of each child outcome (Language skills or Socioemotional skills) on the index built to predict that outcome at each time point. Both variables are standardized. T stands for Treatment group, C for Control group. Estimated coefficients with standard errors in parentheses below.

D Results tables

D.1 Impact analysis

Tables S.9, S.10, S.11, and S.12 below show the results presented in the main manuscript as tables instead of figures. On top of showing the standardized differences between the treatment and control groups and their standard errors, the tables additionally provide the p-values corresponding to the two-sided test of the hypothesis that the difference is null. We present three types of p-values. First, we present the standard p-value calculated without taking into consideration the different hypotheses being tested or the sample size. Second, we present the bootstrapped p-value, which takes into consideration potential distortions in the distribution of the data due to the small sample size. Last, we present the bootstrapped p-values adjusted for multiple-hypothesis testing following the step-down resampling methodology proposed by Romano and Wolf (2016), which controls the family-wise error rate. Since our analyses on the parental speech indices are exploratory, raising the risk of false discoveries, we combined the two indices as well as the Conversational Turns measure together in one family for the family-wise

error rate adjustment in our analyses of parental speech. For the ATE estimations, this means each family consists of three hypotheses. For the CATE estimations that additionally test two subgroups, this means each family consists of six hypotheses. For child outcomes, we grouped in the same family of hypotheses the different measures composing the z-score of language skills in the disaggregated analysis. For the ATE estimations, this means each family consists of two hypotheses. For the CATE estimations that additionally test two subgroups, this means each family consists of four hypotheses. Results are discussed in the main text.

Table S.9: ATEs of interventions on parental speech (random effects model)

	Experiment 1			Experiment 2		
Variable	T-C	p	N	T-C	p	N
Index Language Skills	0.27	*0.12*	962	0.12	*0.12*	295
	(0.18)	*0.15*		(0.07)	*0.04*	
		0.15			*0.04*	
Index Socioemo. Skills	0.11	*0.00*	962	0.15	*0.03*	295
	(0.03)	*0.00*		(0.07)	*0.01*	
		0.00			*0.01*	
Conversational Turns	0.88	*0.00*	861	0.75	*0.00*	192
	(0.18)	*0.00*		(0.26)	*0.00*	
		0.00			*0.00*	

Note: Standardized differences between Treatment and Control groups, with standard errors clustered at the individual level in parentheses below. In the p column, standard unadjusted p-values are shown first, then bootstrapped unadjusted p-values are shown below, and bootstrapped p-values adjusted for multiple-hypothesis testing following Romano and Wolf (2016) are shown last. The family of outcomes contains the two indices as well as the Conversational Turns (3 hypotheses). Two-sided t-tests. N gives the number of observations in each panel regression, which is the number of participants at each time point multiplied by the number of time points (see Figures S.1 and S.2 in the Supplementary Materials).

Table S.10: ATEs of interventions on child outcomes

	Experiment 1			Experiment 2		
Variable	T-C	p	N	T-C	p	N
Language Skills	0.09	*0.82*	258	0.30	*0.06*	59
	(0.39)			(0.16)		
Socioemo. Skills	-0.06	*0.72*	453	0.37	*0.00*	216
	(0.17)			(0.10)		
Language Skills (disaggregated):						
PLS | WM	4.33	*0.00*	919	0.27	*0.10*	59
	(0.47)	*0.00*		(0.16)	*0.10*	
		0.00			*0.12*	
PPVT | ROWPVT	0.02	*0.92*	258	0.33	*0.07*	65
	(0.16)	*0.87*		(0.18)	*0.07*	
		0.87			*0.12*	

Note: Standardized differences between Treatment and Control groups with standard errors in parentheses below. The top half of the table shows results from a random effects model with standard errors clustered at the individual level for panel data and simple OLS regressions with robust standard errors for cross-sectional data. The bottom half of the table shows the same models, separating the two language assessments combined in the 'Language Skills' z-score (PLS and PPVT in Exp. 1 and WM and ROWPVT in Exp. 2 - see Appendix A). In the p column, standard unadjusted p-values are shown first, then bootstrapped unadjusted p-values are shown below, and bootstrapped p-values adjusted for multiple-hypothesis testing following Romano and Wolf (2016) are shown last. P-values are adjusted in the bottom half of the table, for the two language outcomes (2 hypotheses). Two-sided t-tests. N gives the number of observations in each panel regression, which is the number of participants at each time point multiplied by the number of time points (see Figures S.1 and S.2 in the Supplementary Materials).

Table S.11: CATEs of interventions on parental speech (random effects model)

	Experiment 1				Experiment 2			
Variable	T-C (G1)	p	T-C (G2)	p	T-C (G1)	p	T-C (G2)	p
Index Language Skills	0.43	*0.07*	0.20	*0.38*	0.11	*0.16*	0.17	*0.30*
	(0.24)	*0.03*	(0.23)	*0.40*	(0.08)	*0.07*	(0.17)	*0.26*
		0.11		*0.40*		*0.29*		*0.40*
Index Socioemo. Skills	0.14	*0.02*	0.09	*0.01*	0.18	*0.02*	0.06	*0.70*
	(0.06)	*0.00*	(0.04)	*0.00*	(0.08)	*0.01*	(0.16)	*0.64*
		0.02		*0.02*		*0.12*		*0.64*
Conversational Turns	0.69	*0.02*	0.97	*0.00*	0.67	*0.03*	1.03	*0.03*
	(0.29)	*0.00*	(0.23)	*0.00*	(0.31)	*0.01*	(0.48)	*0.06*
		0.02		*0.00*		*0.14*		*0.14*

Note: Standardized differences with standard errors clustered at the individual level in parentheses below. T-C (G1) gives the treatment effect for parents whose highest level of education is a high-school degree or below, T-C (G2) gives the treatment effect for parents whose highest level of education is above a high-school degree. In the p column, standard unadjusted p-values are shown first, then bootstrapped unadjusted p-values are shown below, and bootstrapped p-values adjusted for multiple-hypothesis testing following Romano and Wolf (2016) are shown last. The family of outcomes contains the two indices as well as the Conversational Turns for two subgroups (6 hypotheses). Two-sided t-tests. The number of observations in each regression is identical to the numbers shown in Table S.9.

Table S.12: CATEs of interventions on child outcomes

Variable	Experiment 1				Experiment 2			
	T-C (G1)	p	T-C (G2)	p	T-C (G1)	p	T-C (G2)	p
Language Skills	0.36	*0.60*	0.01	*0.98*	0.34	*0.04*	0.33	*0.40*
	(0.68)	*0.42*	(0.45)	*0.97*	(0.16)	*0.04*	(0.39)	*0.46*
		0.65		*0.97*		*0.15*		*0.46*
Socioemo. Skills	-0.47	*0.15*	0.13	*0.50*	0.55	*0.00*	0.06	*0.68*
	(0.32)	*0.04*	(0.19)	*0.28*	(0.12)	*0.00*	(0.14)	*0.70*
		0.06		*0.28*		*0.00*		*0.70*
Language Skills (disaggregated)								
PLS | WM	4.13	*0.00*	4.24	*0.00*	0.26	*0.14*	0.47	*0.22*
	(0.87)	*0.00*	(0.55)	*0.00*	(0.17)	*0.14*	(0.38)	*0.32*
		0.00		*0.00*		*0.42*		*0.43*
PPVT | ROWPVT	0.14	*0.59*	-0.02	*0.93*	0.47	*0.02*	0.05	*0.92*
	(0.26)	*0.40*	(0.19)	*0.89*	(0.19)	*0.01*	(0.43)	*0.92*
		0.63		*0.89*		*0.14*		*0.92*

Note: Standardized differences between Treatment and Control groups with standard errors in parentheses below. The top half of the table shows results from a random effects model with standard errors clustered at the individual level for panel data and simple OLS regressions with robust standard errors for cross-sectional data. The bottom half of the table shows the same models, separating the two language assessments combined in the 'Language Skills' z-score (PLS and PPVT in Exp. 1 and WM and ROWPVT in Exp. 2 - see Appendix A). In the p column, standard unadjusted p-values are shown first, then bootstrapped unadjusted p-values are shown below, and bootstrapped p-values adjusted for multiple-hypothesis testing following Romano and Wolf (2016) are shown last. P-values are adjusted for the two subgroups being compared in all cases, and additionally in the bottom half of the table, for the two language outcomes (4 hypotheses). Two-sided t-tests. T-C(G1) gives the treatment effect for children whose parents' highest level of education is a high-school degree or below, T-C(G2) gives the treatment effect for children whose parents' highest level of education is above a high-school degree. The number of observations in each regression is identical to the numbers shown in Table S.10.

E Robustness analysis

E.1 Fixed effects model

Tables S.13 and S.14 below are replications of Figures 1 and 3, but they use a fixed effects model instead of a random effects model to estimate the panel regressions, thereby controlling for any time-invariant family-specific sources of variation that could confound the treatment effect. Results are discussed in the manuscript following the discussion of the results from the main specification.

Table S.13: ATEs of interventions on parental speech (fixed effects model)

	Experiment 1			Experiment 2		
Variable	T-C	p	N	T-C	p	N
Index Language Skills	1.33	*0.03*	962	0.10	*0.20*	295
	(0.59)	*0.10*		(0.08)	*0.09*	
		0.10			*0.14*	
Index Socioemo. Skills	0.11	*0.00*	962	0.11	*0.19*	295
	(0.04)	*0.00*		(0.08)	*0.08*	
		0.02			*0.14*	
Conversational Turns	1.28	*0.00*	861	0.78	*0.00*	192
	(0.16)	*0.00*		(0.24)	*0.00*	
		0.00			*0.00*	

Note: Standardized differences between Treatment and Control groups, with standard errors clustered at the individual level in parentheses below. In the p column, standard unadjusted p-values are shown first, then bootstrapped unadjusted p-values are shown below, and bootstrapped p-values adjusted for multiple-hypothesis testing following Romano and Wolf (2016) are shown last. The family of outcomes contains the two indices as well as the Conversational Turns (3 hypotheses). Two-sided t-tests. N gives the number of observations in each panel regression, which is the number of participants at each time point multiplied by the number of time points (see Figures S.1 and S.2 in the Supplementary Materials).

Table S.14: CATEs of interventions on parental speech (fixed effects model)

	Experiment 1				Experiment 2			
Variable	T-C (G1)	p	T-C (G2)	p	T-C (G1)	p	T-C (G2)	p
Index Language Skills	0.73	*0.02*	1.61	*0.06*	0.06	*0.51*	0.26	*0.11*
	(0.30)	*0.01*	(0.85)	*0.21*	(0.09)	*0.39*	(0.16)	*0.15*
		0.12		*0.21*		*0.68*		*0.35*
Index Socioemo. Skills	0.15	*0.02*	0.09	*0.04*	0.13	*0.20*	0.05	*0.73*
	(0.06)	*0.01*	(0.04)	*0.01*	(0.10)	*0.09*	(0.16)	*0.70*
		0.13		*0.17*		*0.35*		*0.70*
Conversational Turns	1.03	*0.00*	1.40	*0.00*	0.76	*0.01*	0.85	*0.10*
	(0.27)	*0.00*	(0.19)	*0.00*	(0.27)	*0.01*	(0.52)	*0.15*
		0.00		*0.00*		*0.11*		*0.35*

Note: Standardized differences with standard errors clustered at the individual level in parentheses below. T-C (G1) gives the treatment effect for parents whose highest level of education is a high-school degree or below, T-C (G2) gives the treatment effect for parents whose highest level of education is above a high-school degree. In the p column, standard unadjusted p-values are shown first, then bootstrapped unadjusted p-values are shown below, and bootstrapped p-values adjusted for multiple-hypothesis testing following Romano and Wolf (2016) are shown last. The family of outcomes contains the two indices as well as the Conversational Turns for two subgroups (6 hypotheses). Two-sided t-tests. The number of observations in each regression is identical to the numbers shown in Table S.13.

E.2 CATE analysis with equalized subgroup sizes

Tables S.15, S.16, and S.17 below are replications of the CATE estimations but they constrain the two subgroups (parents with at most a high-school degree and parents with a degree above

high school) to be the same size, to assess the extent to which differences between the two subgroups could be driven by differences in the size of the two subgroups. In the first experiment, we have 68 parents with at most a high-school degree and 138 parents with a degree above high school at baseline. To equalize the sample sizes, we therefore randomly subsampled 68 parents from the higher-educated group, resulting in 68 parents in each subgroup at baseline. In the second experiment, we have 68 parents with at most a high-school degree and 23 parents with a degree above high school at baseline. To equalize the sample sizes, we therefore randomly subsampled 23 parents from the lower-educated group, resulting in 23 parents in each subgroup at baseline.

Comparing Tables S.15 and S.11, one can note that in general, the tendency in Exp. 1 for the impact on the language skills index to be higher among lower-educated parents and the impact on the socioemotional skills index to be similar across the two groups holds in the bootstrapped sample. In Exp. 2, the difference in impacts on the socioemotional skills index is attenuated when the size of the two subgroups is equalized, with both estimates having p-values far above standard significance levels, suggesting that we cannot reject the hypothesis that these impacts are null. As emphasized in the main manuscript, those results are overall imprecisely estimated, which makes it difficult to draw robust conclusions. The impacts on the conversational turns are more precise and confirm the pattern observed in the full sample: in both experiments, the effect of the intervention tends to be higher among higher-educated families. Table S.16 shows that the patterns observed in the bootstrapped random effects model tend to be also observed in the bootstrapped fixed effects specification. The magnitude of the estimates differ, which is expected given that the fixed effects model only uses the within-individual variation of the data (we refer the reader to the main manuscript, Results section, for additional explanations on each model).

Last, the comparison of Tables S.17 and S.12 suggests that the main findings hold when equalizing the size of the two education subgroups. In Exp. 1, the difference is impacts on socioemotional skills, with lower-educated families experiencing a negative impact while higher-educated families experience a positive but non-significant impact, remains. On the other hand, the impact of the intervention on the PLS is similar in both subgroups (around 4.2-4.4σ). In Exp. 2, the large difference in the impact on socioemotional skills holds (0.62σ, significant at the 5% level, among lower-educated families versus 0.06σ, not significantly different from zero, among higher-educated families). Other estimates are mostly imprecise.

Table S.15: CATEs of interventions on parental speech on bootstrapped sample (random effects model)

Variable	Experiment 1				Experiment 2			
	T-C (G1)	p	T-C (G2)	p	T-C (G1)	p	T-C (G2)	p
Index Language Skills	0.43	*0.07*	-0.03	*0.93*	0.13	*0.18*	0.19	*0.25*
	(0.24)	*0.03*	(0.34)	*0.93*	(0.10)	*0.16*	(0.16)	*0.22*
		0.15		*0.93*		*0.54*		*0.59*
Index Socioemo. Skills	0.14	*0.01*	0.13	*0.01*	0.10	*0.34*	0.06	*0.70*
	(0.06)	*0.00*	(0.05)	*0.00*	(0.10)	*0.27*	(0.16)	*0.64*
		0.04		*0.04*		*0.63*		*0.78*
Conversational Turns	0.66	*0.03*	0.90	*0.00*	0.33	*0.56*	1.04	*0.03*
	(0.30)	*0.00*	(0.30)	*0.00*	(0.58)	*0.56*	(0.49)	*0.05*
		0.08		*0.02*		*0.78*		*0.21*

Note: Standardized differences with standard errors clustered at the individual level in parentheses below. T-C (G1) gives the treatment effect for parents whose highest level of education is a high-school degree or below, T-C (G2) gives the treatment effect for parents whose highest level of education is above a high-school degree. In the p column, standard unadjusted p-values are shown first, then bootstrapped unadjusted p-values are shown below, and bootstrapped p-values adjusted for multiple-hypothesis testing following Romano and Wolf (2016) are shown last. The family of outcomes contains the two indices as well as the Conversational Turns for two subgroups (6 hypotheses). Two-sided t-tests. The number of observations in each regression is identical to the numbers shown in Table S.9.

Table S.16: CATEs of interventions on parental speech on bootstrapped sample (fixed effects model)

Variable	Experiment 1				Experiment 2			
	T-C (G1)	p	T-C (G2)	p	T-C (G1)	p	T-C (G2)	p
Index Language Skills	0.73	*0.02*	0.44	*0.34*	0.09	*0.46*	0.26	*0.12*
	(0.30)	*0.01*	(0.46)	*0.23*	(0.12)	*0.40*	(0.16)	*0.14*
		0.03		*0.23*		*0.75*		*0.43*
Index Socioemo. Skills	0.15	*0.02*	0.12	*0.06*	0.04	*0.77*	0.05	*0.73*
	(0.06)	*0.01*	(0.06)	*0.02*	(0.13)	*0.72*	(0.16)	*0.70*
		0.03		*0.06*		*0.91*		*0.91*
Conversational Turns	1.03	*0.00*	1.25	*0.00*	0.65	*0.16*	0.85	*0.11*
	(0.27)	*0.00*	(0.23)	*0.00*	(0.46)	*0.14*	(0.52)	*0.14*
		0.00		*0.00*		*0.43*		*0.43*

Note: Standardized differences with standard errors clustered at the individual level in parentheses below. T-C (G1) gives the treatment effect for parents whose highest level of education is a high-school degree or below, T-C (G2) gives the treatment effect for parents whose highest level of education is above a high-school degree. In the p column, standard unadjusted p-values are shown first, then bootstrapped unadjusted p-values are shown below, and bootstrapped p-values adjusted for multiple-hypothesis testing following Romano and Wolf (2016) are shown last. The family of outcomes contains the two indices as well as the Conversational Turns for two subgroups (6 hypotheses). Two-sided t-tests. The number of observations in each regression is identical to the numbers shown in Table S.13.

Table S.17: CATEs of interventions on child outcomes on bootstrapped sample

Variable	Experiment 1				Experiment 2			
	T-C (G1)	p	T-C (G2)	p	T-C (G1)	p	T-C (G2)	p
Language Skills	0.37	*0.59*	-0.23	*0.74*	0.18	*0.53*	0.33	*0.42*
	(0.69)	*0.40*	(0.69)	*0.60*	(0.29)	*0.54*	(0.41)	*0.47*
		0.64		*0.64*		*0.71*		*0.71*
Socioemo. Skills	-0.47	*0.15*	0.22	*0.34*	0.62	*0.02*	0.06	*0.69*
	(0.33)	*0.04*	(0.23)	*0.14*	(0.26)	*0.02*	(0.14)	*0.70*
		0.07		*0.14*		*0.04*		*0.70*
Language Skills (disaggregated)								
PLS | WM	4.37	*0.00*	4.19	*0.00*	0.04	*0.90*	0.47	*0.24*
	(0.86)	*0.00*	(0.80)	*0.00*	(0.32)	*0.90*	(0.39)	*0.31*
		0.00		*0.00*		*0.99*		*0.59*
PPVT | ROWPVT	0.14	*0.59*	-0.08	*0.77*	0.47	*0.12*	0.05	*0.92*
	(0.26)	*0.40*	(0.28)	*0.64*	(0.29)	*0.14*	(0.44)	*0.92*
		0.64		*0.64*		*0.45*		*0.99*

Note: Standardized differences between Treatment and Control groups with standard errors in parentheses below. The top half of the table shows results from a random effects model with standard errors clustered at the individual level for panel data and simple OLS regressions with robust standard errors for cross-sectional data. The bottom half of the table shows the same models, separating the two language assessments combined in the 'Language Skills' z-score (PLS and PPVT in Exp. 1 and WM and ROWPVT in Exp. 2 - see Appendix A). In the p column, standard unadjusted p-values are shown first, then bootstrapped unadjusted p-values are shown below, and bootstrapped p-values adjusted for multiple-hypothesis testing following Romano and Wolf (2016) are shown last. P-values are adjusted for the two subgroups being compared in all cases, and additionally in the bottom half of the table, for the two language outcomes (4 hypotheses). Two-sided t-tests. T-C(G1) gives the treatment effect for children whose parents' highest level of education is a high-school degree or below, T-C(G2) gives the treatment effect for children whose parents' highest level of education is above a high-school degree. The number of observations in each regression is identical to the numbers shown in Table S.10.

F Additional analyses

Table S.18: Evolution of disparities in ASQ-SE in T vs C group (Exp. 2)

	Treatment group		Control group	
Time point	G1-G2	N	G1-G2	N
24-30 months	-0.22	42	-0.24	40
	(0.28)		(0.44)	
30-36 months	0.36	32	-0.48	36
	(0.15)		(0.33)	
36-42 months	-0.13	32	-0.64	34
	(0.25)		(0.32)	

Note: Differences in standardized socioemotional scores between children whose parents' highest level of education is a high-school degree or below (G1) and children whose parents' highest level of education is above a high-school degree (G2). Standard errors in parentheses below. N gives the number of observations in each regression.

SI References

Bainter, S. A., T. G. McCauley, T. Wager, and E. A. R. Losin (2020). Improving practices for selecting a subset of important predictors in psychology: An application to predicting pain. *Advances in Methods and Practices in Psychological Science 3*(1), 66–80.

Cristia, A., M. Lavechin, C. Scaff, M. Soderstrom, C. Rowland, O. Räsänen, J. Bunce, and E. Bergelson (2021). A thorough evaluation of the language environment analysis (LENA) system. *Behavior Research Methods 53*, 467–486.

Er, M. B. and I. B. Aydilek (2019). Music emotion recognition by using chroma spectrogram and deep visual features. *International Journal of Computational Intelligence Systems 12*(2), 1622–1634.

Kong, Q., Y. Cao, T. Iqbal, Y. Wang, W. Wang, and M. D. Plumbley (2020). PANNs: Large-scale pretrained audio neural networks for audio pattern recognition. *IEEE/ACM Transactions on Audio, Speech, and Language Processing 28*, 2880–2894.

Krajewski, J., A. Batliner, and M. Golz (2009). Acoustic sleepiness detection: Framework and validation of a speech-adapted pattern recognition approach. *Behavior Research Methods 41*(3), 795–804.

List, J. A. (2020). Non est disputandum de generalizability? a glimpse into the external validity trial. Technical report, National Bureau of Economic Research.

Low, L.-S. A., N. C. Maddage, M. Lech, L. Sheeber, and N. Allen (2010). Influence of acoustic low-level descriptors in the detection of clinical depression in adolescents. In *2010 IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 5154–5157. IEEE.

MacWhinney, B. (2000). *The CHILDES project: Tools for analyzing talk: Volume I: Transcription format and programs, volume II: The database*. MIT Press One Rogers Street, Cambridge, MA 02142-1209, USA.

Mauch, M. and S. Dixon (2014). pyin: A fundamental frequency estimator using probabilistic threshold distributions. In *2014 ieee international conference on acoustics, speech and signal processing (icassp)*, pp. 659–663. IEEE.

McFee, B., C. Raffel, D. Liang, D. P. Ellis, M. McVicar, E. Battenberg, and O. Nieto (2015). librosa: Audio and music signal analysis in Python. In *Proceedings of the 14th python in science conference*, Volume 8, pp. 18–25.

Nogueira, W., T. Rode, and A. Büchner (2016). Spectral contrast enhancement improves speech intelligibility in noise for cochlear implants. *The Journal of the Acoustical Society of America 139*(2), 728–739.

Rejaibi, E., A. Komaty, F. Meriaudeau, S. Agrebi, and A. Othmani (2022). MFCC-based recurrent neural network for automatic clinical depression recognition and assessment from speech. *Biomedical Signal Processing and Control 71*, 103107.

Romano, J. P. and M. Wolf (2016). Efficient computation of adjusted p-values for resampling-based stepdown multiple testing. *Statistics & Probability Letters 113*, 38–40.

Sin, W. L., B. Y. Chang, X. Ma, and A. Horner (2019). *The acoustics features and their relationship to the emotional characteristics of rain sound effects.* Ph. D. thesis, Hong Kong University of Science and Technology.

Suskind, D. L., K. R. Leffel, E. Graf, M. W. Hernandez, E. A. Gunderson, S. G. Sapolich, E. Suskind, L. Leininger, S. Goldin-Meadow, and S. C. Levine (2016). A parent-directed language intervention for children of low socioeconomic status: A randomized controlled pilot study. *Journal of Child Language 43*(2), 366–406.

Taguchi, T., H. Tachikawa, K. Nemoto, M. Suzuki, T. Nagano, R. Tachibana, M. Nishimura, and T. Arai (2018). Major depressive disorder discrimination using vocal acoustic features. *Journal of Affective Disorders 225*, 214–220.

VanDam, M., A. S. Warlaumont, E. Bergelson, A. Cristia, M. Soderstrom, P. De Palma, and B. MacWhinney (2016). HomeBank: An online repository of daylong child-centered audio recordings. In *Seminars in speech and language*, Volume 37, pp. 128–142. Thieme Medical Publishers.

Weninger, F., F. Eyben, B. W. Schuller, M. Mortillaro, and K. R. Scherer (2013). On the acoustics of emotion in audio: what speech, music, and sound have in common. *Frontiers in Psychology 4*, 292.

Whittingslow, D. C., J. Zia, S. Gharehbaghi, T. Gergely, L. A. Ponder, S. Prahalad, and O. T. Inan (2020). Knee acoustic emissions as a digital biomarker of disease status in juvenile idiopathic arthritis. *Frontiers in Digital Health 2*, 571839.

Zhang, Y., N. Charoenphakdee, Z. Wu, and M. Sugiyama (2020). Learning from aggregate observations. *Advances in Neural Information Processing Systems 33*, 7993–8005.