**Supplemental Material**

*S.1: Time-Varying Control Covariates Measured at 7, 15, and 24 months of age*

Family income-to-needs level was also measured, based on primary-caregiver self-report of household income, divided by the poverty threshold for family size. Child prescription or over-the-counter medication use was dummy-coded (0= no, 1 = yes). Time of day was measured on a 24-hour clock. The passage of time was measured in years, centered on 7 months.

Time-Invariant Control Covariates:

The following demographic information at 2 months of age: education level (years); marital status (0 = non-married, 1 = married), education level (years).

At 2 months of age, primary-caregiver functional reading ability was included as a control and measured using the *KFAST* literacy screener (Kaufman & Kaufman, 1994) a standardized measure of functional literacy. Higher scores indicate better reading skills. Reliability and validity of the *KFAST* are well established (Flanagan et al., 1997).

Primary caregiver depression and anxiety levels were based on self-reports on the *Brief Symptom Inventory-18* (BSI-18; Derogatis, 2000), a short, highly sensitive, self-report screening index for psychological distress. Each scale was based on the mean of 6 items and standardized per Derogatis (2000) to create a t-score. Higher scores indicate more problematic symptomology. Published internal consistency reliability (α) for the Depression and Anxiety scales are .84 and .79, respectively.

Two-month child covariates include child’s birth weight (kg), gender (1 = male), race (1 = African American), and observation site (1 = NC).

Mean values of each of the respective time-varying predictors over time were calculated and included as time-invariant control covariates.

*Selection Predictors included in Generalized Propensity Score (GPS) Model*

All control time-invariant control covariates listed above were included in the GPS Model

Additional Measures at 2 months (Maternal Self-Report):

Mother’s first child is target child (0 -1), Mothers age at first child (years); Mother has asthma (0 – 1); Mother had ADD (0-1); Mother had learning disability(0-1); Mother ever repeated a grade (0-1); Mother has High Blood Pressure (0-1); Mother is overweight (0-1); Mother has diabetes (0-1); Mother has had miscarriage (0-1); Mother smoked during pregnancy with target child (0-1); Mother had happy pregnancy (0-1); Mother has emotional problems (0-1); Mother has family problems (0-1); Mother weekly work hours (hours); Mother receives WIC (0-1); Mother receives TANF; Mother receives SNAP (food stamps; 0-1); Mother receives childcare subsidy (0-1); Mother breast feeds (0-1); Hours TV is on in the house (hours);

Other:

Child Head Circumference (cm)

*S.2: Description of Generalized Propensity Score Approach*

The logic behind inversed probability weights is informed by the *Potential Outcomes* framework of causality (Rubin, 1974, 2005). For each child *i* there presumed to be is a set of outcomes Yi(t) for a given level of treatment (t ∈ T). For instance, in the context of more typical binary treatment regimens, treatment is defined as, t ∈ (0,1), with the counterfactual reflecting one’s outcome conditional on receiving the opposite of one’s actual treatment status (i.e., treatment or control). In other words, if child *i* was treated, what is his/her outcome when she is not treated? For a non-treated child, one flips the question. In the context of continuous a treatment, *T* represents the interval t0,t1, with the counterfactual indicating one’s outcome had he/she simultaneously received treatment one unit lower than his/her actual treatment level. Of course, one individual cannot simultaneously receive the treatment and control. As such, this “missing data” problem (Rubin, 2005) opens up the estimate of the average causal effect, E[T(1) – T(0)], to bias because the effect may be confounded by unobserved factors that exist between children experiencing different levels of the treatment.

Notably, as shown by Rosenbaum and Rubin (1983), given the assumption that children with identical combinations of observed covariate values are functionally equivalent with respect to a unit difference in the continuous treatment—that is, there is no unobserved confounding of treatment (*Y*(*t*)⊥*T* |*X* ∀ *t*∈*T)—*one can recover an estimate of the average causal effect of the treatment.

A common approach to accomplish this is to specify an omnibus metric of background characteristics—a propensity score—which can, in turn, be used to ensure that one’s treatment effects are based on children who, aside from the treatment difference, are functionally equivalent (i.e., similar propensity for treatment level). For binary treatments, propensity scores are typically estimated using the predicted probabilities of treatment, via a logistic regression model, in which treatment is regressed on a large set of potential pre-treatment confounders. Hirano and Imbens (2004) showed that a similar propensity score can be estimated for continuous treatments, under the assumption that treatment is normally distributed, with a mean and a variance (*Ti*∣∣*Xi*∼*N* (*β*0+*β*′1*Xi*, *σ*2). Specifically, treatment *T* is regressed on a vector of confounders. The probability of attaining a given treatment level is obtained from a the normal probability density function of the treatment, conditional on the confounders,

$ϕ\left(t\right)= \frac{1}{2π\hat{σ}^{2}}e ^{\frac{-(t-\hat{t})^{2}}{2σ^{2}}}$,

where $ϕ\left(t\right)$ is n.d.f , *t;* $π$ is a constant (i.e., 3.14); $\hat{t}$ i s the predicted treatment value; *t* is the observed treatment value, and $\hat{σ^{2}}$ is the residual standard error from the linear regression model. Considered another way, the propensity score is the probability of the standardized residual, given a normal distribution—or the probability of a given level of treatment, conditional on the confounders.

The inverse of the propensity score is estimated to reflect the probability of receiving a treatment one unit different than the treatment he/she actually received. As such, the IPWTs essentially create a pseudo-population in which those at a given level of the treatment approximate those receiving a unit lower of the treatment. That is, they down-weight units that are over-represented for a given level of the treatment and up-weight units that are under-represented for that given level of treatment. Following Robins (2000), because GPSs with continuous treatments can lead to unrealistically high probabilities, IPWTs are typically stabilized by the dividing GPS estimates derived from an “empty” linear regression model by those from the GPS estimates derived from the “full” model with the selection covariates.

Notably, the following assumptions are implicit in our model: (1) a normally distributed treatment, such that the conditional probability of receiving a given level of treatment can be estimated using a normal probability density function; (2) no unmeasured confounds exist outside of those modeled (exchangability); (3) causal estimates are based on variation in treatments level that exist within the multivariate distribution of confounders—sometimes called the “region of common support” or the positivity assumption. In other words, the effects of maternal sensitivity cannot be based on maternal sensitivity data that extrapolates to non-existent children, with respect to the confounds; (4) correct specifications for both the selection and substantive models.

To estimate our GPS scores, we first imputed missing data across five datasets using the Amelia II package (Honaker , King, & Blackwell, 2011) for R (R Development Core Team, 2005). We estimated GPSs by regressing mean maternal sensitivity on 45 potential pre-treatment confounds. These selection-model predictors were obtained from a larger pool of all demographic, maternal, and infancy variables available at 2 months of age.

OLS models were fitted iteratively (Proc Reg; SAS), transforming and interacting variables as needed in order to achieve balance of the covariates across levels parental sensitivity (see below). As above, we used the standardized residuals from an “empty” model and from the respective “final” conditional models to obtain estimate the respective probabilities of the outcome, given an n.d.f—dividing the latter by the former to create the stabilized inverse probability weights (see Robins et al., 2000). The entire series of models and estimations were conducted for each of the five imputed datasets and the IPWTs were averaged the weights across imputations. Similar to prior work (Sampson, Sharkey, & Raudenbush, 2008) we trimmed extremely high weights (> 99th percentile) and imputed 99thpercentile values to reduce their undue leverage.

 As shown in Figure S.1, the weights removed the vast majority of the imbalance. After weighting, formerly moderate to strong associations between maternal sensitivity and the selection variables were rendered largely orthogonal. All weighted correlations were well below .20, with the majority of them below .05. Selection variables showing modest imbalance were included as covariates in the multi-level models testing our substantive questions.

Figure S.1. *Correlations between selection variables (i.e., confounders) and Mean Maternal Sensitivity levels across 7, 15, 24 months, pre- and post-IPWT weighting. Smaller correlations reflect balance.*

References

Derogatis L (2000). *Brief symptom inventory. Vol. 18*. Minneapolis: NCS Pearson, Inc.

Kaufman, A., S. & Kaufman, N.,L. (1994). *KFAST: Kaufman functional academic skills test*. Toronto, Canada: Psycan.

Flanagan, D. P., McGrew, K. S., Abramowitz, E., Lehner, L., Untiedt, S., Berger, D., & Armstrong, H. (1997). Improvement in academic screening instruments? A concurrent validity investigation of the K-FAST, MBA, and WRAT-3. *Journal of Psychoeducational Assessment*, *15*, 99-112. doi: 10.1177/073428299701500201

Hirano, K., & Imbens, G. (2004). The propensity score with continuous treatments. In aA. Gelman & X.L. Meng ( Eds.). *Applied Bayesian modeling and causal inference from incomplete-data perspectives.* New York: Wiley.

Honaker J., King, G., & Blackwell M. (2009). *Amelia II: A program for missing data*. Cambridge, MA: Harvard University.

Imai, K., & Van Dyk, D. A. (2004). Causal inference with general treatment regimes. *Journal of the American Statistical Association*, *99,854-866*. doi:10.1198/016214504000001187

Kirschbaum, C., & Hellhammer, D. H. (1989). Salivary cortisol in psychobiological research: An overview. *Neuropsychobiology, 223*, 150–169. doi: 10.1159/000118611

Robins, J. M. (2000). Marginal structural models versus structural nested models as tools for causal inference. *Statistical models in epidemiology, the environment, and clinical trials*. Springer New York.

Rosenbaum, P. R., & Rubin, D. B. (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika*, *70*), 41-55. *doi: 10.1093/biomet/70.1.41*

Rubin, D. B. (1974). Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of educational Psychology*, *66*, 688-701. doi: [10.1037/h0037350](http://psycnet.apa.org/doi/10.1037/h0037350%22%20%5Ct%20%22_blank)

Rubin, D. B. (2005). Causal inference using potential outcomes. *Journal of the American Statistical Association*, *100,322-331*. doi: 10.1198/016214504000001880

Sampson, R. J., Sharkey, P., & Raudenbush, S. W. (2008). Durable effects of concentrated disadvantage on verbal ability among African-American children. *Proceedings of the National Academy of Sciences*, *105*, 845-852. doi: 10.1073/pnas.0710189104