Effects of an Early Childhood Intervention on Adult Mental Health

Caitlin Kearns

January 19, 2016

1 Introduction

A substantial literature examines the effect of early childhood programs on long-run outcomes. A number of these studies use data from small-scale randomized interventions, namely the Carolina Abecedarian Project (ABC) and the Perry Preschool Program (PPP). Anderson (2008) finds that both the ABC and PPP have significant effects on an index of adult outcomes, but only for females. Heckman, Moon, Pinto, Savelyev and Yavitz (2010) find significant effects of the PPP on both male and female outcomes, including educational attainment, employment, and criminal activity, though the type and timing of the most significant effects differ by sex. Heckman, Pinto and Savelyev (2013) extend these results by modeling the mechanisms through which the PPP improved adult outcomes, and find that most of the effect is explained by a reduction in externalizing (aggressive, dishonest, or antisocial) behaviors. Campbell, Conti, Heckman, Moon, Pinto, Pungello and Pan (2014) and Conti, Heckman and Pinto (2015) find that the ABC had positive effects on healthy behaviors as well as metabolic and cardiovascular risk factors, particularly for men.

In contrast to physical health, long-run effects on mental health have not been studied within the context of early childhood programs. Like educational attainment, employment, and

physical health behaviors, mental health has important implications for both productivity and well-being. Given the effects of the PPP on childhood externalizing behavior, it is possible that preschool directly affects psychological traits which carry forward into adulthood. In addition, treated individuals may experience less psychological distress in adulthood as a result of improved economic outcomes and better physical health. While I do not attempt to model the production function for mental health, I use the randomized nature of the ABC to estimate causal effects on adult psychological characteristics. I find little evidence of significant effects on mental health at age 21, though there is some evidence of reduced hostility for females.

2 Data

The ABC was an intensive preschool program which provided educational, nutritional, and health care services to treated children from age zero to five. Children were selected at birth or shortly after, with eligibility based on a high risk index measuring family disadvantage and risk of developmental delay, and randomly assigned to treatment and control groups (Campbell, Conti, Heckman, Moon, Pinto, Pungello and Pan 2014).

I use data from the initial ABC data collection and age 21 follow-up study. My primary source of outcomes is the Subject Brief Symptom Inventory (BSI). The BSI is a 53-item questionnaire which measures psychological risk factors on a five point scale across nine dimensions, including anxiety, depression, and hostility. A large body of research has examined the validity and sensitivity of the BSI (Rath and Fox 2011).

Before proceeding to the analysis, it is important to clarify which aspects of mental health the

data are measuring. The BSI is designed as an indicator of emotional-behavioral functioning and psychological distress, not as a diagnostic tool (Rath and Fox 2011). While most items of the BSI correspond to internal experiences (e.g. feelings of guilt, spells of terror or panic), some items are closely tied to behaviors and actions (e.g. getting into frequent arguments, having to double check things). As an indicator of mental health, therefore, the BSI reflects general well-being and daily functioning rather than indicating a particular psychiatric condition, although certain subscales (such as depression, obsessive-compulsivity, and psychoticism) may signal a diagnosable mental illness. Given that mental health is a broad term, other studies may adopt a different definition and employ different sources of data when examining mental health.

3 Research design

While the ABC is a randomized controlled trial, the small scale of the program, large number of outcomes, and sample attrition have the potential to complicate treatment effect estimation.[1] For small scale inference and multiple hypothesis testing, I follow the procedures outlined by Anderson (2008). I first normalize the 53 BSI items and compute subscales corresponding to the nine emotional-behavioral dimensions, as well as a Global Severity Index (GSI); using subscales both reduces the dimensionality of the testing problem and aids in interpretation.[2] For each of these ten outcomes, I compute both naïve p-values based on Huber-White standard errors and nonparametric permutation p-values. To account for

[1] In contrast to the PPP, adjustments do not need to be made for deviations from ideal randomization in the treatment and control assignments (Conti, Heckman and Pinto 2015).

[2] The normalized subscale for each observation is equivalent to the efficient GLS estimator from a regression of the corresponding BSI items on a constant.

multiple-hypothesis testing, I recompute p-values using the free stepdown procedure, which provides strong familywise error rate (FWER) control. Most of the literature outlined above uses the Romano and Wolf stepdown procedure, which requires somewhat weaker assumptions than the free stepdown procedure; I therefore also test for significance at the 5% and 10% levels using the Romano and Wolf procedure, but do not report p-values. I conduct all tests for females and males separately, as is standard in the literature.

The third complication, sample attrition, is of fairly small magnitude in the ABC sample: the original sample has 111 subjects, while the subject Brief Symptom Inventory has 104 (6% attrition). However, some items are missing for certain individuals, which results in missing subscale values, particularly for the GSI (which is a composite of the nine subscale scores). Anderson (2008) does not adjust for attrition; Campbell et al. (2014) and Conti et al. (2015) adjust for attrition using inverse probability weighting, but their sample of biomedical outcomes (collected in subjects' mid-30s) suffers from substantially higher attrition than the age 21 surveys. I therefore treat the problem of attrition as second order and focus on adjusting for the small sample size and multiple hypothesis testing.

4 Descriptive statistics

Tables 1 and 2 show the initial covariate balance for females and males. Since the ABC intervention began in infancy, I only consider variables determined at birth (other than entry age). Age of entry differs significantly between the treatment and control groups for females and males, while mother's high school completion differs significantly for females only. Following Heckman, Moon, Pinto, Savelyev and Yavitz (2010), I adjust for these imbalances

Table 1: Covariate balance for females (Fraction treated = 0.475)

Variable	N	Mean, treated	Mean, control	Difference	p-value
Entry age, weeks	59	8.036	4.484	3.552	0.000942
Mother's age less than 18	58	0.630	0.581	0.0490	0.709
Mother's WAIS score	59	85.50	82.10	3.403	0.255
Mother graduated high school	55	0.560	0.267	0.293	0.0286
High risk index	58	20.56	23	-2.444	0.107
Father living in home	59	0.750	0.677	0.0726	0.545
Apgar score	51	9.087	8.857	0.230	0.466
Gestation age, weeks	52	39.62	40.50	-0.875	0.135

All outcomes other than age of entry are measured at birth. The Wechsler Adult Intelligence Scale (WAIS) is a test of cognitive ability. The high risk index was used to determine program eligibility. Differences are estimated from OLS regressions of each variable on the treatment indicator. p-values are based on robust Huber-White standard errors.

Table 2: Covariate balance for males (Fraction treated = 0.558)

Variable	N	Mean, treated	Mean, control	Difference	p-value
Entry age, weeks	51	9.393	5.826	3.567	0.0156
Mother's age less than 18	51	0.607	0.565	0.0419	0.768
Mother's WAIS score	52	85.41	87.04	-1.630	0.566
Mother graduated high school	51	0.321	0.304	0.0171	0.898
High risk index	51	19.43	18.96	0.472	0.716
Father living in home	52	0.724	0.652	0.0720	0.588
Apgar score	46	8.708	8.955	-0.246	0.363
Gestation age, weeks	50	39.11	39.22	-0.106	0.859

All outcomes other than age of entry are measured at birth. The Wechsler Adult Intelligence Scale (WAIS) is a test of cognitive ability. The high risk index was used to determine program eligibility. Differences are estimated from OLS regressions of each variable on the treatment indicator. p-values are based on robust Huber-White standard errors.

by conditioning on these two covariates in all regressions.[3]

5 Results

Table 3: Normalized outcomes for females

Outcome	N	Mean	SD	Effect	Naive p-value	Permutation p-value	FWER p-value
Somatization, BSI	51	-0.0242	0.715	0.191	0.417	0.421	0.465
Obsessive-compulsive, BSI	51	-0.0246	0.676	0.360	0.133	0.139	0.380
Interpersonal sensitivity, BSI	50	-0.0289	0.673	0.422	0.0608	0.0608	0.250
Depression, BSI	51	-0.0204	0.576	0.444	0.0373	0.0421	0.208
Anxiety, BSI	51	-0.0220	0.598	0.320	0.0675	0.0661	0.250
Hostility, BSI	51	-0.0201	0.720	0.657	0.00631	0.00730	0.0517
Phobic anxiety, BSI	49	-0.0154	0.575	0.142	0.431	0.451	0.465
Paranoid ideation, BSI	48	0.00616	0.617	0.194	0.298	0.305	0.465
Psychoticism, BSI	50	-0.0274	0.602	0.408	0.0362	0.0376	0.206
Global Severity Index, BSI	46	0.0123	0.730	0.393	0.115	0.121	0.377

Effects estimated from OLS regressions of each outcome on the treatment indicator, age of entry fixed effects, and a dummy indicating that the mother completed high school. Naïve p-values are based on heteroscedasticity robust Huber-White standard errors; permutation p-values and FWER p-values are computed according to Anderson (2008).

Main results: Tables 3 and 4 show the results for females and males. All variables are normalized and scaled so that a higher value indicates better mental health, i.e. a lower incidence of the corresponding risk factor. The treatment effect is the coefficient from a regression of the outcome variable on the treatment indicator, age of entry fixed effects, and a dummy for mother's high school completion. In both samples, the naïve and permutation p-values are very similar.

Estimated treatment effects are generally larger and more significant for females than for males, consistent with Anderson (2008). For females, all treatment effects are positive, and

[3]Anderson (2008) does not condition on any covariates. Campbell et al. (2014) and Conti et al. (2015) condition on additional covariates from the ABC biomedical data, but I am not able to match these variables to the age 21 data I am using.

Table 4: Normalized outcomes for males

Outcome	N	Mean	SD	Effect	Naive p-value	Permutation p-value	FWER p-value
Somatization, BSI	48	-0.0107	0.590	0.200	0.335	0.343	0.805
Obsessive-compulsive, BSI	50	-0.0104	0.659	0.290	0.277	0.292	0.774
Interpersonal sensitivity, BSI	50	0.0365	0.722	0.0899	0.713	0.712	0.885
Depression, BSI	49	0.0569	0.642	0.379	0.152	0.157	0.615
Anxiety, BSI	51	-1.05e-08	0.587	0.445	0.122	0.126	0.558
Hostility, BSI	47	0.0163	0.708	0.361	0.212	0.214	0.710
Phobic anxiety, BSI	50	-0.00520	0.534	0.0170	0.933	0.932	0.932
Paranoid ideation, BSI	51	3.69e-09	0.690	-0.165	0.531	0.536	0.852
Psychoticism, BSI	49	-0.00641	0.613	0.288	0.225	0.230	0.713
Global Severity Index, BSI	41	0.0333	0.732	0.279	0.385	0.408	0.805

Effects estimated from OLS regressions of each outcome on the treatment indicator, age of entry fixed effects, and a dummy indicating that the mother completed high school. Naïve p-values are based on heteroscedasticity robust Huber-White standard errors; permutation p-values and FWER p-values are computed according to Anderson (2008).

several appear significant in the absence of multiple hypothesis testing adjustments (depression, hostility, and psychoticism at the 5% level; interpersonal sensitivity and anxiety at the 10% level). When considering FWER p-values computed using the free stepdown procedure, the effect on the hostility[4] subscale score is large in magnitude (close to one standard deviation) and significant at the 10% level; the treatment effect for all other outcomes is insignificant.[5] For males, all treatment effects are insignificant, even without accounting for the small sample size and multiple hypothesis testing. The Romano and Wolf stepdown procedure produces the same results, except that the effect on hostility for females is significant at the 5% level.

Robustness: The results in tables 3 and 4 represent the preferred specification. Considering

[4]From the BSI introductory report: "The hostility dimension is organized around three categories of hostile behaviour: thoughts, feelings, and actions. Typical experiences cover feelings of annoyance and irritability, urges to break things, frequent arguments and uncontrollable outbursts of temper" (Derogatis and Melisaratos 1983).

[5]More precisely, the stepdown procedure rejects at the 10% level that all treatment effects are equal to zero, but fails to reject that all treatment effects other than hostility are equal to zero.

raw[6] rather than normalized subscale scores does not meaningfully change the results: the effect on hostility for females is of similar magnitude (measured in standard deviations), though somewhat less significant (FWER p-value = 0.09). When no covariates are included in the regressions, the effect on the normalized hostility score for females is substantially smaller (0.389) and insignificant (FWER p-value = 0.18).

Childhood hostility: Given that the educational component of the ABC included intensive small group interactions between children and between children and teachers, it is plausible that the program improved interpersonal skills and consequently reduced hostile thoughts and behaviors (Campbell et al. 2014). If so, it is natural to consider whether this effect was already apparent in childhood. The original ABC data contain measures of hostility[7] corresponding to kindergarten, first grade, and second grade. The estimated effect for females is negative in grades one and two, though I fail to reject the null of no effect on all three measures using the free stepdown procedure. This result is in contrast to Heckman, Pinto and Savelyev (2013), who find significant differences in externalizing behavior (which includes aggression or hostility as a component) between the PPP treatment and control groups for ages six to nine, for both females and males.[8]

[6]The raw subscale score is simply the average of the subscale component items.

[7]The childhood hostility measure is from the Classroom Behavior Inventory (CBI), a teacher questionnaire. Hostility is rated on a scale from 3 to 15.

[8]Their measure is derived from the Pupil Behavior Inventory (PBI) and the Ypsilanti Rating Scale (YRS), not the CBI. The PBI and YRS are also teacher reported.

6 Conclusion

I find some evidence that the ABC substantially reduced adulthood feelings of hostility in females; however, there is little other indication of significant long run effects on mental and emotional health, particularly for males. The effect on female hostility is marginally significant and is not necessarily robust to changes in the set of conditioning covariates. In addition, a large reduction in hostility in adulthood is difficult to reconcile with no effect on hostility in childhood.

A possible explanation for the lack of an effect on males is the timing of the survey: there is some evidence that effects for females manifest earlier than for males. Heckman, Moon, Pinto, Savelyev and Yavitz (2010) find significant effects of the PPP on criminal activity and economic outcomes for males aged 27 to 40, while significant effects for females are found in early adulthood (ages 19 to 27). Likewise, the larger health effects for males found by Campbell et al. (2014) are derived from data collected in subjects' mid-30s. If mental health measures had also been collected in this more recent data sweep, it is possible that more significant effects would have been found for males.

When interpreting the results, it is also important to note that the BSI is a limited measure of mental health. The survey is relatively brief (each subscale is comprised of only four to seven items), and individual evaluations of symptom severity may be subjective. Also, since all items are self-reported, subjects may choose to under-report symptoms of distress (Rath and Fox 2011). A parent-reported version of the BSI is available at age 21; parent responses may also be problematic, however, since it is not clear how much contact parents and children had

at the time of the survey. It is possible that alternate indicators of mental health would produce more or less significant results.

As a further step, it may be worthwhile to use inverse probability weighting to determine whether attrition and missing values are affecting the results. A formal analysis of the mechanisms through which the ABC affects adult mental health would also be a natural next step, though it may not be possible to follow the procedure of Heckman, Pinto and Savelyev (2013) given the variables available in the ABC data. Given the range of adult outcomes available for the ABC, it could also be informative to examine the relationship between mental and physical health, or between mental health and behaviors such as criminal activity; such an analysis would acquire additional modeling assumptions beyond the randomization assumption I exploit here.

References

Anderson, Michael, "Multiple Inference and Gender Differences in the Effects of Early Intervention: A Reevaluation of the Abecedarian, Perry Preschool, and Early Training Projects," *Journal of the American Statistical Association*, 2008, *103*, 1481–1495.

Campbell, Frances, Gabriella Conti, James Heckman, Seong Hyeok Moon, Rodrigo Pinto, Elizabeth Pungello, and Yi Pan, "Early childhood investments substantially boost adult health," *Science*, 2014, *343*, 1478–1485.

Conti, Gabriella, James Heckman, and Rodrigo Pinto, "The Effects of Two Influential Early Childhood Interventions on Health and Healthy Behaviors," *NBER Working Paper 21454*, 2015.

Derogatis, Leonard R. and Nick Melisaratos, "The Brief Symptom Inventory: An introductory report," *Psychological Medicine*, 1983, *13*, 595–605.

Heckman, James, Rodrigo Pinto, and Peter Savelyev, "Understanding the mechanisms through which an influential early childhood program boosted adult outcomes," *American Economic Review*, 2013, *103*, 2052–2086.

— , Seong Hyeok Moon, Rodrigo Pinto, Peter Savelyev, and Adam Yavitz, "Analyzing Social Experiments as Implemented: A Reexamination of the Evidence from the HighScope Perry Preschool Program," *Quantitative Economics*, 2010, *1*, 1–46.

Rath, Joseph F. and Lisa M. Fox, "Brief Symptom Inventory," in Jeffrey S. Kreutzer, John DeLuca, and Bruce Caplan, eds., *Encyclopedia of Clinical Neuropsychology*, New York: Springer, 2011, pp. 449–451.

7 Data appendix

7.1 Data

ABC: The original ABC data and age 21 follow up data are available from ICPSR (studies 4091 and 32262, respectively). A description of the recruitment, randomization, and treatment procedures for the ABC is available in section A of the supplementary materials for Campbell et al. (2014). For a comparison of the main characteristics of the ABC and PPP, see Conti et al. (2015), Table 1.

Mental health outcomes: In the initial analysis, I considered outcomes from three subject surveys: the Brief Symptom Inventory (BSI); the What I am Like survey, known more widely as the Self-Perception Profile for Adults (SPP); and the risk taking survey.

- BSI: The BSI has 53 items corresponding to nine subscales. Each item is rated on a scale from zero (no occurrence of the symptom) to four (extreme occurence). Each raw subscale score is a simple average of the items included in the subscale. The Global Severity Index (GSI) is a composite of all subscale scores. I consider all nine subscales and the GSI.

- SPP: The SPP has 50 items corresponding to 12 subscales, each reflecting an aspect of self-image or perceived competence. Each item is rated on a scale from one to four. Each raw subscale score is a simple average of the items included in the subscale. I considered four subscales: sociability, nurturance, intimate relationships, and global self-worth.

- Risk taking: The risk taking survey has four items relating to suicide risk. I considered an indicator for whether the subject considered suicide in the past 12 months.

For the BSI and SPP, I construct subscale scores and the GSI using the following steps based on Anderson (2008), section 3.2.1:

1. Match individual items to subscales using Derogatis and Melisaratos (1983) for the BSI and Messer and Harter (2012) for the SPP. To verify the classification, I replicated the raw subscale scores included in the surveys. I was able to exactly replicate all subscale scores, but not the GSI; the correlation between the survey GSI and my computed GSI variable (a simple average of the nine subscale scores) is greater than 0.99.

2. Recode the underlying items so that a higher value indicates a more desirable mental health outcome.

3. Recompute raw subscale scores and raw GSI using the recoded items.

4. Standardize the underlying items by demeaning and dividing by the control group standard deviation.

5. For each subscale k with N_k corresponding items, and each subject i, compute the normalized subscale score $z_{ik} = [\mathbf{1}_k' \mathbf{W}_k \mathbf{1}_k][\mathbf{1}_k' \mathbf{W}_k \mathbf{y}_{ik}]^{-1}$, where $\mathbf{y}_{ik}$ is a $N_k \times 1$ vector of standardized subscale items, $\mathbf{W}_k$ is the efficient $N_k \times N_k$ weighting matrix (the inverse of the covariance matrix for the items corresponding to subscale k), and $\mathbf{1}_k$ is a $N_k \times 1$ vector of ones.

7.2 Research design

For the procedure used to compute permutation p-values, see Anderson (2008), section 3.1. For details of the free stepdown procedure, see Anderson (2008), section 3.2.2. For the

Romano and Wolf stepdown procedure, see Romano and Wolf (2005) and section H of the supplementary materials for Campbell et al. (2014). For all tests, I set the number of iterations to 10,000, which is less than the number recommended by Anderson (2008)[9] but greater than the number used by Heckman, Pinto and Savelyev (2013). While the free stepdown p-values change somewhat between runs, the results of the hypothesis tests remain stable.

7.3 Results

The results for the BSI subscales are similar when including the SPP and risk taking variables in the set of hypotheses. Since the SPP subscales and suicide indicator are highly insignificant, I omit them from the set of hypotheses tested in the preferred specification presented in section 5.

[9]Anderson (2008) recommends 100,000 iterations, which would further stabilize the p-values between runs. This number of iterations was not feasible for computational reasons.

References

Anderson, Michael, "Multiple Inference and Gender Differences in the Effects of Early Intervention: A Reevaluation of the Abecedarian, Perry Preschool, and Early Training Projects," *Journal of the American Statistical Association*, 2008, *103*, 1481–1495.

Campbell, Frances and Elizabeth Pungello, "Carolina Abecedarian Project (ABC) and the Carolina Approach to Responsive Education (CARE), Age 21 Follow Up Study, 1993 - 2003." ICPSR32262-v1. Ann Arbor, MI: Inter-university Consortium for Political and Social Research[distributor], 2014-01-31. http://doi.org/10.3886/ICPSR32262.v1.

— , Gabriella Conti, James Heckman, Seong Hyeok Moon, Rodrigo Pinto, Elizabeth Pungello, and Yi Pan, "Early childhood investments substantially boost adult health," *Science*, 2014, *343*, 1478–1485.

Conti, Gabriella, James Heckman, and Rodrigo Pinto, "The Effects of Two Influential Early Childhood Interventions on Health and Healthy Behaviors," *NBER Working Paper 21454*, 2015.

Derogatis, Leonard R. and Nick Melisaratos, "The Brief Symptom Inventory: An introductory report," *Psychological Medicine*, 1983, *13*, 595–605.

Heckman, James, Rodrigo Pinto, and Peter Savelyev, "Understanding the mechanisms through which an influential early childhood program boosted adult outcomes," *American Economic Review*, 2013, *103*, 2052–2086.

Messer, Bonnie and Susan Harter, "The Self-Perception Profile for Adults: Manual and Questionnaires," 2012.

Ramey, Craig T., James J. Gallagher, Frances A. Campbell, Barbara H. Wasik, and Joseph J. Sparling, "Carolina Abecedarian Project and the Carolina Approach to Responsive Education (CARE), 1972-1992." ICPSR04091-v1. Ann Arbor, MI: Inter-university Consortium for Political and Social Research [distributor], 2004. http://doi.org/10.3886/ICPSR04091.v1.

Romano, Joseph P. and Michael Wolf, "Exact and Approximate Stepdown Methods for Multiple Hypothesis Testing," *Journal of the American Statistical Association*, 2005, *100*, 94–108.